REPORT DOCUMENTATION PAGE

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE 31 May 96	3. REPORT TYPE AND DATES COVERED

4. TITLE AND SUBTITLE
Methodology To Determine Homology and Clustering as Applied To Intronic Regions Of Regulatory Cancer Genes and Non-Regulatory Genes

5. FUNDING NUMBERS

6. AUTHOR(S)
Deborah Leigh Hall

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Colorado

8. PERFORMING ORGANIZATION REPORT NUMBER
96- 023D

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
DEPARTMENT OF THE AIR FORCE
AFIT/CI
2950 P STEET, BLDG 125
WRIGHT-PATTERSON AFB OH 45433-7765

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

19961212 058

14. SUBJECT TERMS			15. NUMBER OF PAGES 155
			16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT

GENERAL INSTRUCTIONS FOR COMPLETING SF 298

The Report Documentation Page (RDP) is used in announcing and cataloging reports. It is important that this information be consistent with the rest of the report, particularly the cover and title page. Instructions for filling in each block of the form follow. It is important to *stay within the lines* to meet *optical scanning requirements*.

Block 1. Agency Use Only *(Leave blank)*.

Block 2. Report Date. Full publication date including day, month, and year, if available (e.g. 1 Jan 88). Must cite at least the year.

Block 3. Type of Report and Dates Covered. State whether report is interim, final, etc. If applicable, enter inclusive report dates (e.g. 10 Jun 87 - 30 Jun 88).

Block 4. Title and Subtitle. A title is taken from the part of the report that provides the most meaningful and complete information. When a report is prepared in more than one volume, repeat the primary title, add volume number, and include subtitle for the specific volume. On classified documents enter the title classification in parentheses.

Block 5. Funding Numbers. To include contract and grant numbers; may include program element number(s), project number(s), task number(s), and work unit number(s). Use the following labels:

C	-	Contract	**PR** -	Project
G	-	Grant	**TA** -	Task
PE	-	Program Element	**WU** -	Work Unit Accession No.

Block 6. Author(s). Name(s) of person(s) responsible for writing the report, performing the research, or credited with the content of the report. If editor or compiler, this should follow the name(s).

Block 7. Performing Organization Name(s) and Address(es). Self-explanatory.

Block 8. Performing Organization Report Number. Enter the unique alphanumeric report number(s) assigned by the organization performing the report.

Block 9. Sponsoring/Monitoring Agency Name(s) and Address(es). Self-explanatory.

Block 10. Sponsoring/Monitoring Agency Report Number. *(If known)*

Block 11. Supplementary Notes. Enter information not included elsewhere such as: Prepared in cooperation with...; Trans. of...; To be published in.... When a report is revised, include a statement whether the new report supersedes or supplements the older report.

Block 12a. Distribution/Availability Statement. Denotes public availability or limitations. Cite any availability to the public. Enter additional limitations or special markings in all capitals (e.g. NOFORN, REL, ITAR).

DOD - See DoDD 5230.24, "Distribution Statements on Technical Documents."
DOE - See authorities.
NASA - See Handbook NHB 2200.2.
NTIS - Leave blank.

Block 12b. Distribution Code.

DOD - Leave blank.
DOE - Enter DOE distribution categories from the Standard Distribution for Unclassified Scientific and Technical Reports.
NASA - Leave blank.
NTIS - Leave blank.

Block 13. Abstract. Include a brief *(Maximum 200 words)* factual summary of the most significant information contained in the report.

Block 14. Subject Terms. Keywords or phrases identifying major subjects in the report.

Block 15. Number of Pages. Enter the total number of pages.

Block 16. Price Code. Enter appropriate price code *(NTIS only)*.

Blocks 17. - 19. Security Classifications. Self-explanatory. Enter U.S. Security Classification in accordance with U.S. Security Regulations (i.e., UNCLASSIFIED). If form contains classified information, stamp classification on the top and bottom of the page.

Block 20. Limitation of Abstract. This block must be completed to assign a limitation to the abstract. Enter either UL (unlimited) or SAR (same as report). An entry in this block is necessary if the abstract is to be limited. If blank, the abstract is assumed to be unlimited.

METHODOLOGY TO DETERMINE HOMOLOGY AND CLUSTERING

AS APPLIED TO INTRONIC REGIONS OF

REGULATORY CANCER GENES AND NON-REGULATORY GENES

by

DEBORAH LEIGH HALL

B.A., Russell Sage College, 1982

M.S., University of Southern California, 1983

M.S., Creighton University, 1988

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Preventive Medicine and Biometrics

1996

This thesis for the Doctor of Philosophy degree by

Deborah Leigh Hall

has been approved for the
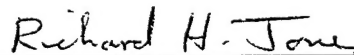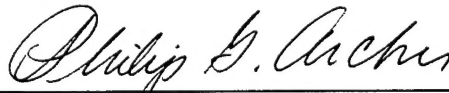
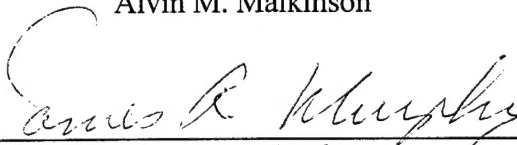Department of Preventive Medicine and Biometrics

by

_____
Karen Kafadar

_____
Richard H. Jones

_____
Philip G. Archer

_____
Alvin M. Malkinson

_____
James R. Murphy

Date ___May 31, 1996___

Hall, Deborah Leigh (Ph.D., Analytic Health Sciences/Biometrics)

Methodology to Determine Homology and Clustering as Applied to Intronic Regions of

Regulatory Cancer Genes and Non-regulatory Genes.

Thesis directed by Associate Professor Karen Kafadar.

This thesis considers the problem of statistically evaluating the closeness of multiple genes based on their DNA intronic regions.
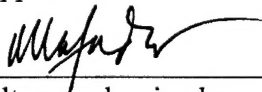
The purpose of intronic regions in a gene is unknown. Recent research suggests that, in cancer related genes, intronic regions may play a role in regulating disease susceptibility. To investigate whether the intronic features of cancer related genes differ from non-regulatory genes, a collection of oncogenes, tumor suppressor genes, and non-regulatory genes involved in enzyme metabolism are analyzed. A statistical methodology is employed to determine whether features of these genes' intronic regions will result in clustering by regulatory group.

This is the first comparison of intronic attributes for 33 *homo sapien* genes. Attributes analyzed include mean of intronic log lengths, standard deviations of intronic log lengths and number of intronic regions in a gene. The analysis also includes the available ordered DNA data from these intronic regions. The process begins with creation of first-order Markov transition counts for all intronic regions in each gene. A hypothesis testing approach employs these counts to determine whether the order of the nucleotides in each intronic region is random or whether this order shows statistical evidence of a first-order Markov chain.

A test of homology on the Markov transition counts from comparable intronic regions of two genes is performed. The average of the Chi-square test statistics for all intronic pairs between two genes creates a measure of homology that depicts the overall closeness between each gene pair. Results for all gene pairs are combined into a distance matrix, demonstrating that results of a statistical hypothesis test can be used as a distance metric. A hierarchical cluster analysis is performed based on this distance matrix.

Several significant biological results were observed. Sixty-seven percent of intronic regions in the data set were first-order Markov processes, providing evidence that not all intronic regions are random DNA sequences. Tumor suppressors clustered based on homology results. Additionally, oncogenes and non-regulatory genes clustered based on number of introns, and based on a combination of number of introns and means of the intronic log lengths in a gene.

The form and content of this abstract are approved. I recommend its publication.

Signed _____

Faculty member in charge of thesis

DEDICATION

I dedicate this dissertation to my family. First, to my brother Marshal for giving me an interest in cancer research. Without him, my choices in life and my research interests would have taken an entirely different path.

Second, to my Dad for always making me put the correct scientific units on my math problems so that I would never forget that the real reason I was doing math problems was scientific discovery.

Third, to my Mom and sister Kathleen for their support during the past year. My sister's electronic mails citing the adventures of my nieces and nephew were a welcome source of comic relief.

Lastly, and most importantly to my husband Paul who is truly my "soul mate". I could not have done this without him always being there to tell me that things would be okay. He is an example of organization and self-discipline that although I will never match, I deeply respect and learn from daily. Words cannot measure how grateful I am to him for all that he has done to help me through not only the past three years, but through life in general.

ACKNOWLEDGMENTS

CONTENTS

FIGURES

CHAPTER I

INTRODUCTION

<u>Purpose of the Study</u>

The primary goal of this thesis is the development of a statistical methodology to describe, quantify and compare currently known and available intronic regions from oncogenes, tumor suppressor genes, and non-regulatory genes found in the process of enzyme metabolism. A test of homology on Markov transition counts from comparable intronic regions is performed on all possible pairs of genes. The average of the Chi-square test statistics for all intronic pairs between each gene pair creates a measure of homology that depicts the overall closeness between the gene pair. On a more global scale, the issue of whether intronic regions of regulatory and non-regulatory genes cluster based on degree of regulatory function is investigated. Hierarchical cluster analysis is used throughout this thesis to determine gene groupings.

Chapter II details previous biological, statistical and computer contributions to this research area. Chapter III describes the data collection process and provides basic information about the genes in the data set.

Chapter IV contains the first published comparison of intronic attributes for the *homo sapien* genes examined in the data set. Available intronic features include: number of introns in a gene, mean of the intronic log lengths in a gene, and the standard deviation of intronic log lengths in a given gene.

Chapter V answers the question of whether intronic sequence data in different regulatory groups are totally random or show evidence of a first order Markov process.

Numerous articles state that a Markov process is present in DNA sequences (see Chapter II). Mostly focused on the coding regions, a few references suggest that Markov structure may also hold for introns. There remains a prevalent belief in the biological community that introns do not exhibit evidence of a Markov process. By use of a hypothesis testing approach, this thesis show that the 67 percent of introns in the data set do exhibit evidence of a Markov process.

Assuming Markov processes, a statistical hypothesis test is applied to determine whether comparable introns from each gene have homologous Markov transition matrices in Chapter VI. The resulting Chi-square test statistics are averaged for each gene and placed into a distance matrix that is used in the clustering process. This provides evidence that results of a statistical test can be used in the creation of a distance matrix. This is also the first systematic look at intronic regions both within a gene and among genes.

Results of all cluster analyses in this thesis are examined to ascertain whether there is any consistent evidence of clustering by gene regulatory function. The success of this methodology suggests that a similar approach can be taken when the data set includes a gene of unknown regulatory status.

Scope of the Study

This thesis addresses statistical issues related only to the human nuclear intronic regions in oncogenes, tumor suppressor genes, and non-regulatory genes (involved in the process of enzyme metabolism). Because the evolutionary conservation of intronic regions between species is known to be relatively weak, it is not prudent to generalize these results to other organisms. Similarly, this study draws no conclusions concerning

2

highly regulatory genes related to diseases other than cancer, nor does it provide conclusions concerning non-regulatory genes involved in human maintenance processes other than enzyme metabolism. Lastly, as all introns analyzed in this study are nuclear introns, generalizations to other intronic groups (i.e., groups I, II and III introns and transfer ribonucleic acid (tRNA) introns) should not be made. These groups are known to be different (Hickson, 1989).

The methodology in this thesis seeks a global measure of closeness between two genes that does not utilize traditional gene alignment and scoring based on percentage match. There are times when local and global alignments with matching are necessary and preferable; e.g., multiple alignment of protein products (where the biological function of the protein products is relatively well explored), searches for transcription factors, searches for oligonucleotides, and single alignment searches of highly conserved coding regions. This thesis is not intended denigrate work accomplished in these areas, but offers an alternative approach to current methods when drawing conclusions about the relationship between two genes based on the entire intronic data available. The homology test used in this thesis is based on Markov processes rather than on aligned matches that provide a composite score.

Although this research is preliminary, it is hoped that the conclusions will help to determine possible avenues for future biological laboratory research. Results observed may change as more intronic gene sequence data become available and the methodology can be enhanced to incorporate more data.

CHAPTER II

REVIEW OF THE LITERATURE

Biological Review

Gene Structure

Genes are made of deoxyribonucleic acid (DNA), which codes for the production of polypeptide chains that are joined in specific orders to produce proteins (Friedman, *et al.*, 1992). DNA molecules exist primarily in the cell nucleus and consist of two complementary chains that twist to form a double helix. Each chain is composed of four nucleotides that contain a deoxyribose sugar, a phosphate, and a pyrimidine or purine base. Purine bases are adenine (A) and guanine (G); pyrimidine bases are cytosine (C) and thymine (T). Thymine occurs only in DNA and is replaced by Uracil (U) in ribonucleic acid (RNA). Ribonucleic acid is similar to DNA except that it is composed of a ribose sugar and is found in the cytoplasm of the cell as well as the nucleus.

Role of Intronic Regions

In 1977, it was discovered that eukaryotic genes are interrupted (Lewin,1994). The interruption of eukaryotic genes means there are huge blocks of DNA that do not code for proteins. Although the existence of these interrupted regions is now known, the main focus of most genetics research is the DNA coding regions, equivalently called exonic regions, exons, or the translated portion of DNA (Lewin, 1994). Yet, less than 3 percent of DNA carries code that produces proteins. The purpose of the remaining 97 percent of the code is currently unknown. Explanations for these regions in the literature are sparse and often conflicting (Wills, 1991).

Three types of non-coding regions are found in eukaryotic DNA. These are intronic regions, untranslated regions and intergenic spacers. For the purpose of this thesis, intronic regions of nuclear genes are the only non-coding regions I will discuss and analyze.

Intronic regions, or introns, are non-coding segments of a gene that are transcribed but then removed from the transcript by splicing together the coding regions (exons) on either side of them. Once removed, they are quickly degraded in the nucleus (King and Stansfield, 1990). Transcription is the process by which the genetic information on the sense strand of DNA is transcribed into a complementary messenger RNA (Wills, 1991). With very rare exceptions, the sequence *GU* represents the beginning of an intron and the sequence *AG* designates the end of an intron (Goodman, 1994 and Lewin, 1994). Further studies have been done to determine if there is more of a pattern than simply a particular dinucleotide at each end. The results have not been consistent (Penotti, 1991 and Gelfand, 1989).

Several theories provide possible explanations for the purpose of intronic DNA. These range from the theory that introns are meaningless evolutionary relics i.e., "junk DNA" (Flam, 1994) to the theory that intronic regions somehow control gene regulation (Wills, 1994). As specific examples of intronic regions controlling regulation, it was proposed that intronic structures of cancer-related genes such as K-*ras* and H-*ras* regulate susceptibility to cancer (You *et al.*, 1992, Hashimoto-Gotoh *et al.*, 1988, Malkinson and You, 1994, Telliez *et al.*, 1995). Laboratory experiments that intentionally splice intronic regions out of the gene before transcription have demonstrated an increase or decrease in protein production (Bordonaro, 1994 and Bossu, 1994). Although these examples

5

support the premise that intronic regions do have some sort of regulatory purpose, the possible mechanism behind this process is currently unknown.

The lack of consensus over intronic purpose has unfortunately resulted in a large hole in the literature when it comes to quantifying and classifying known facts about these non-coding regions. Another major reason for this lack of information relates to the purpose behind the Human Genome Organization (HUGO). Until very recently, HUGO did not solicit the collection and analysis of intronic information. The primary goals of HUGO give the highest priority to mapping and determination of the complete human DNA sequence (Pearson and Soll, 1991). However, a gene's total coding sequence defines a complete sequence. Published DNA code need not include the intronic regions to be designated as complete in the molecular genetics literature.

At the present time, the sequence author decides whether to enter intronic information into GenBank (national repository for DNA sequence data sponsored by the National Center for Biotechnology Information (NCBI)). Many authors do not enter non-coding information because, unless they have a computer with a digitizer, they have to enter the code into GenBank manually via INTERNET. It is tedious to enter these regions that can easily be 100,000 base pairs (bp) long (Lewin, 1994). This, plus the belief that introns may serve no useful purpose, leads many researchers to disregard entering intronic sequence code into GenBank. The sporadic nature of intronic sequence data entry into GenBank and the difficulty of manipulating GenBank for exploratory data searches result in little research concerning introns.

Only a few literary sources quantify intronic regions. One fact consistently reported in the literature is that the number of introns in a gene can vary greatly across genes.

Similarly, the lengths of these non-coding regions vary greatly both between genes and within a given gene (Lewin, 1994). However, a statistical definition of "varies greatly" and simple information like mean intronic length and variance of intronic length for a given human gene is frequently unknown. The most basic descriptive statistics are unavailable. Questions of interest like whether genes with very small numbers of introns (or very large numbers of introns) tend to be alike in function remain unanswered. A paper by Hawkins (1988) is frequently referenced by the biological community as a sources for determination of intron length in various parts of genes for vertebrates, insects, plants and fungi. The differences between phyla are depicted in Table 2.1.

Table 2.1: Average Length and Number of Introns Collected by Hawkins

	5' Introns	Internal Introns	3' Introns
Vertebrates	1811(91)	1127(1941)	681(25)
Insects	5507(19)	622(210)	
Fungi		86(126)	
Plants		249(200)	

Average lengths are provided in bp. Numbers of introns collected are provided in parentheses. 5' introns (i.e., introns at the beginning of a gene) are introns wholly within the 5' non-coding region. 3' introns (i.e., introns at the end of a gene) are introns within the 3' untranslated portion of the gene. Not all intronic information was present for every gene (Hawkins, 1988).

Although Hawkins attempted to do a very thorough job, the paper suffers from the limitations of information known at that time. First, the sample size is extremely small. Second, in 1988 the biological estimates of true lengths for long introns were not available. This created a degree of selection bias in his results. Third, data for humans are contained within the vertebrates class. If there are differences moving up the ladder

to the higher organisms, there may be differences within the vertebrate phylum as well. This means the group of greatest biological importance to us is somewhat obscured by the remainder of the vertebrate data. Lastly and most importantly, Hawkins does not differentiate between intronic lengths in regulatory genes versus non-regulatory genes; hence, potential differences between these two groups are confounded by the lack of stratification.

In 1988, Hawkin's study was a time consuming task based on current information. In 1996, this paper should be re-evaluated and replaced with more modern information. Although still far from perfect, information in this thesis will reflect a more accurate picture than was possible ten years ago.

Rationale for the Selection of Regulatory and Non-regulatory Genes

Regulatory genes are genes whose primary function is to control the rate of synthesis of the products of other distant genes (King and Stansfield, 1990). Oncogenes are normal genes involved in the regulation of cell growth, but can lead to neoplasia (abnormal tissue growth) when mutated, overexpressed or amplified (i.e., a portion of the DNA sequence is replicated) When this occurs that they induce uncontrolled cell proliferation (Friedman *et al.*, 1990). Tumor suppressor genes are normal genes whose function is to prevent the development of neoplasia, but when mutated the ability to prevent this development is lost (Friedman, 1990). Non-regulatory genes involved in enzyme metabolism are genes whose primary function involves the physical and chemical processes by which living cells produce and maintain themselves, and by which energy is made available for use by the organism (King and Stansfield, 1990).

Degree of mRNA stability appears to play a role in regulating the level of gene expression during cell growth and differentiation (Schiavi, 1992). In mammalian cells, mRNA half-lives range from as short as minutes to as long as days. Among the shortest-lived mRNAs are the *FOS* and *MYC* proto-oncogene mRNAs. These have protein half-lives of approximately 20-30 minutes and mRNA half-lives of approximately 30 minutes (Hargrove, 1989). The short half-lives of these mRNAs appear to play a critical role in the normal function of these genes. The quick deregulation of *FOS* or *MYC* mRNA may somehow contribute to oncogenesis (Schiavi, 1992). As a comparison, one of the non-regulatory genes chosen for this thesis, glyceraldehyde-3-phosphate dehydrogenase, has a protein half-life of 75-130 hours and a mRNA half-life of 8 hours (Hargrove, 1989), a half-life much longer than the oncogenes' half-lives.

The issue of mRNA stability and gene regulation raises several questions of interest. How different are the half-lives of oncogenes and tumor suppressor genes from non-regulatory genes? Does this difference have anything to do with the known physical features that depict that gene (e.g., length of the gene, number of intronic regions, etc.) or with the intronic DNA sequence code within that particular gene? Is there a combined effect from number of introns, the means and standard deviations of the intornic log lengths, and the DNA sequence code? Do genes tend to cluster based on degree of regulatory function? Answers to these questions are unexplored in the literature. The gene groupings in this thesis seek to provide answers to the above questions.

Cluster Analysis and DNA Sequence Analysis

Cluster analysis is the formal study of methods and algorithms for organizing data objectively into groups, or collections of groups (Hirtle, 1995). Because groupings of data are often of interest in DNA research, cluster analysis has been used widely. Some of the oldest uses of clustering and classification are connected to the development of lineage among plants and animals (Hartigan, 1975). Aristotle was the first to use classification on 500 plants and animals based on a comparison system he developed. Linnaeus developed the present classification system of plants and animals in 1753. The tree structure (dendrogram) that is often associated with standard clustering procedures was used by taxonomists after Darwin developed his evolutionary theories. Not surprisingly, the majority of the DNA statistical literature concerns the use of cluster analysis as related to evolutionary issues (King *et al.*, 1979, Meyer, 1991, Ladunga, 1992). Another area that has received much recent attention is clustering groups of cells based on various cellular properties (Mukwaya and Welch, 1989, Musser *et al.*, 1990, Garrels *et al.*, 1990, van der Mei *et al.*, 1991).

Unfortunately, there has not been much work using classification and clustering in grouping DNA sequence data. No papers use cluster analysis on intronic sequences. The few papers relating to sequence analysis generally used only one method of clustering to investigate a particular biological issue (Rowe, *et al.*, 1984, Authn and Fieldes, 1992, Bickam *et al.*, 1995,). Not much progress has been made to incorporate what is known statistically about the data into the analysis. Additionally, most researchers do not

question the merits of the particular clustering approach that they use; they use whatever clustering algorithm to which they have access, regardless of how well it suits their data.

Markov Processes and DNA Sequence Analysis

The use of Markov processes in DNA research has also been widely explored. Some articles focus on the use of Markov processes in nucleotide sequence analysis, but the majority of research is applied to the creation of models to simulate evolutionary events. Since evolutionary issues are beyond the scope of this paper, this review focuses on only those methods concerning nucleotide sequence analysis.

To perform a statistical analysis of DNA, one desires an adequate mathematical model. After the first sequencing of DNA, it was observed that a model of independent random sequences was probably inappropriate. Numerous models were proposed (Gelfand *et al.*, 1992), including: stationary and non-stationary processes, Markov models, hidden Markov chains, mixture transition distribution models, block models, and finite automata. Despite a substantial amount of research, a completely adequate model for the description of DNA/protein sequences is still an open problem. No existing algorithm has been able to generate a long sequence (> 10,000 bp) that can pass as a naturally occurring DNA sequence. Markov chains are probably the most consistent and widely chosen of the possible models, although their use has not been without controversy. Neither increasing the Markov chain order nor taking into account biased frequencies of nucleotide runs, explains all the anomalies found in genetic code.

Older examples of DNA sequence analysis using Markov chains date back to the 1970's (Fuchs, 1980). In early analyses, researchers believed they would find a Markov order that best fit all DNA sequences. Gatlin (1972) examined first-order Markov chain

11

models. Hasegawa and Yano (1975), and Figueroa *et al.* (1978) examined second-order Markov chain models. Garden (1980) provided evidence against the previous belief of a possible "best" order when he concluded that different orders of Markov chains optimally fit different DNA sequences. In his analysis, a third-order model best fit a ϕX174 bacteriophage sequence, a second-order best fit both the early and late regions of the SV40 bacteriophage, and the nucleotides in the MS2 gene (partial sequence) were independent.

Considering these results, Garden developed a modeling methodology to determine the optimal order of a given Markov chain by applying Tong's (1975) procedure. He applied this methodology to a Markov chain with a finite number of states. The basis for Tong's procedure is the Akaike information criterion (AIC). If $L(k)$ is the likelihood corresponding to the kth order model, the number of independent parameters for the model is $\tau(k) = 4^{k+1} - 4^k$. The criterion for selection of model order is based on the statistic $R(k)$ as follows (Fuchs, 1980):

$$\{k: R(k) = -2\ln L(k) + 2\tau(k) \text{ is minimized}\}$$

Another method for determining the Markov chain order (based on Hoel, 1954) used a likelihood ratio test. In this case, $_k\eta_{k+1}$ is the - 2 ln likelihood ratio such that

$$_k\eta_{k+1} = -2[\ln L(k) - \ln L(k+1)].$$

If $_k\eta_{k+1}$ exceeds the critical value of a chi-square distribution with degrees of freedom given by

$$df(k+1|k) = \tau(k+1) - \tau(k),$$

the difference in fit between the kth and the $(k+1)^{st}$ order models is statistically significant. Fuchs (1980) compared the Tong procedure and the likelihood ratio test procedure on 7 complete DNA sequences (all bacteriophages, viruses and plasmids). He hoped to obtain concordance between the two methods but instead found discrepancies (i.e., each method picked a different optimal order for each of the seven genes). He concluded that the order of the model selected by the $R(k)$ criterion was directly proportional to the number of nucleotides in the sequence (i.e., model order increased as the length of the DNA sequence increased). A possible explanation is that longer DNA sequences are indeed different. Alternatively, and more likely, the longer the sequence, the higher the order of the model. In short sequences, the frequencies of the nucleotide groups are small. When this occurs, the power of the goodness of fit test diminishes and there is a tendency to accept models of lower order.

This raises the issue of whether modeling approaches common in the field of time series analysis are appropriate for the analysis of a long DNA sequence. Fuchs recommended the addition of an analysis of residuals when investigating the order of a sequence. This allows investigation of differences between observed and expected frequencies in groups of nucleotides, plus examination of model order.

Recent DNA sequence analyses using Markov chains follow Fuch's recommendations. Many analyses determine a kth order Markov transition matrix that provides frequencies for words of length k. The observed frequencies are compared against the expected frequencies to provide statistical evidence of the chosen model. Examples of this methodology (i.e., the same basic methods applied to different DNA sequences) are Phillips *et al.*, 1987, Barrai *et al.*, 1990, Lewis *et al.*, 1995. These authors

13

searched for a plausible biological conclusion based on an observed pattern in a DNA region after laboratory sequencing.

Only one paper incorporates the areas of Markov chains and DNA homology (Lake, 1994). The paper was written for the purpose of evolutionary comparisons, but a portion of the methodology could be applied to DNA nucleotide homology. The paper involves paralinear distances. Paralinear distance measures the distance between two sequences. These measures are easy to calculate, valid for nucleotide sequences, and gaps (positions where one or more nucleotides are missing in a strand of DNA) may be included. The paralinear distance between two sequences i and j is

$$d_{ij} = -log_e \frac{det \ J_{ij}}{(det \ D_i)^{1/2}(det \ D_j)^{1/2}} .$$

To calculate the paralinear distance between sequences i and j, one first computes J_{ij}, the joint probability matrix. For example, sequence 1 and sequence 2 are both 10 bp long. Each sequence consists of only nucleotides C and T. The two sequences are as follows:

Sequence 1: CTTCCCTCTC

Sequence 2: TTTCTTCCTC

J_{ij} is represented by

	Sequence 2		
Sequence 1		C	T
C	3	3	6
T	1	3	4
	4	6	

Once J_{ij} is computed, diagonal matrices D_i and D_j can be constructed from the J_{ij} matrix

where $D_1 = \begin{bmatrix} 6 & 0 \\ 0 & 4 \end{bmatrix}$ (i.e., diag (row totals) and $D_2 = \begin{bmatrix} 4 & 0 \\ 0 & 6 \end{bmatrix}$ (i.e., diag (columns totals).

14

The quantity $d_{ij} = 2 \ln 2 = 1.386$ determines the degree of similarity between the two sequences.

This method has some inconsistencies. Even though it claims to be based on a Markov chain, the frequency matrix is based simply on common alignments between the two sequences, not a transition matrix. Also, there is no underlying methodology to use this paralinear distance in a hypothesis testing approach, so there is no way to judge the value of d_{ij}. Thus, this method does not help determine whether two sequences are statistically similar or different.

Recently, applied researchers have moved away from classical Markov chain methodology and toward linguistic approaches (Sakakibara *et al*, 1994, Mantagna *et al.*, 1994, Dong and Searls, 1994) and algorithm based scoring methods (Arratia *et al.*, 1988, Mott *et al.*, 1989). The mathematical and statistical theorists who have continued with the Markov chain methodology often suffer because their methodology is not easily applied to large amounts of DNA sequence data (Gentleman and Mullin, 1989, Kleffe and Langbecker, 1990, Kleffe and Borodovsky, 1992, Kelly, 1994).

While the Markov chain methodology is still an underlying basis for statistical DNA research, the focus has shifted away from achieving the best theoretical methods and instead has moved toward methods that can better handle large amounts of DNA sequence data. In some instances, this means a weakening of the underlying assumptions behind the model, but this trade-off is necessary in the development of computer algorithms for extraordinarily long sequences.

Computer Review

Computer Algorithms and Intronic Regions

Since the mid-1970's when rapid sequencing of DNA became available, researchers developed computer programs to handle various kinds of DNA analyses. With the creation of HUGO in 1988, the number of these programs has proliferated. Some of these programs can process intronic sequences but there is no evidence in the literature of a computer algorithm designed specifically to aid in our understanding of intronic regions. Rather, there is a whole class of algorithms (used primarily for DNA mapping) that seeks to identify introns so that they can be removed from the data, keeping only the coding regions. Additionally, there is another whole family of models (called multiple alignment models) that cannot handle the analysis of introns. Only single alignment algorithms (algorithms that determine homology) are presently useful in comparing intronic regions. But even these have some constraints. These three groups of models and how they affect our knowledge of introns are discussed below.

Models Incorporating Features of Intronic Regions to Aid in DNA Mapping

There has been a great amount of DNA mapping in the past 10 years; e.g., Andrzej Konopka (Biolingua Labs), Gary Stormo (University of Colorado at Boulder) and Eric Snyder (Sequana Therapeutics, Inc.) have worked in the areas of mapping and non-coding sequence analysis (Konopka and Smythers, 1987, Konopka et al., 1987, Stormo and Snyder, 1993). Most of this research is aimed at identifying features of intronic regions, not as possible identifiers of susceptibility to cancer, but for subsequent removal of these intronic regions in long sequences of unknown DNA data. Now that rapid sequencing techniques are available, researchers piece together long sequences of DNA

16

from a designated region of a chromosome. It is estimated that by using these processes, the whole human genome will be sequenced in the next five years (Johnston, 1996). These sequenced chromosomal regions often contain numerous unknown genes. Researchers try to determine features that distinguish introns and exons, and then write computer algorithms to flag the exons and introns based on these unique features. These algorithms assess, with some degree of probability, the location of the numerous genes along a chromosomal DNA sequence (Snyder, 1995). From these gene sequences, they determine the protein for which the gene codes. The combination of computers and laboratory work is efficient and saves considerable time and money so that more genes can be found and mapped in a short time period.

These algorithms sometimes include observations about introns. Unfortunately, since it is not the main focus of the publication, information about introns may be fragmentary. The specific details of each computer program used to detect differences between exon and intron boundaries relates to DNA mapping and is beyond the scope of this thesis.

Multiple Alignment Algorithms and Intronic Regions

This group of computer programs takes a group of sequences that are similar in function and creates a multiple alignment of them. This alignment is stored in the database as a matrix called a "block" and future unknown sequences can then be matched against these blocks. The purpose of searching a database of blocks is to detect repeated domains and find distinct cross-family relationships that had been missed in searches of sequence databases. By searching the data in block form, a researcher has a better chance of finding a short sequence because the blocks focus on only the areas of high conservation and thus eliminate "noise" in the sequences (Henikoff and Henikoff, 1994).

17

Some algorithms used for this method are PROTOMAT (finds the best path for reconstructing all the sequences once the DNA has been cut), PATMAT (takes the DNA portion of interest and searches the known DNA data looking for areas of high conservation (i.e., relatively little change during an evolutionary period of time) and BLOCKSORT (sorts the sequence blocks based on comparision with an existing sequence and scores them). Some block searchers can be accessed from an electronic mail server.

There are two basic approaches to multiple sequence alignment (Boguski, 1992). Global alignment attempts to match a group of sequences along their entire lengths. It is most applicable to sequences that are short and approximately equal in length. Local alignment methods are better suited to longer sequences that vary considerably in length and share only small isolated regions with little or no sequence conservation, or for sequences that are known to contain internal repeats.

It might seem as though these types of programs could be extended easily to intronic sequences or even to complete genes, but there are limitations which prohibit this. This methodology depends upon the fact that there is a known function associated with a given block and a relatively high degree of conservation between the family members of a block. Neither of these features is generally true of introns. A second problem is that, although one of these algorithms (PATMAT) can handle a sequence of 5000 bp, most of them are designed for much shorter sequences. Some programs cannot exceed even 200 bp in length before they start "bogging down". This is a problem for introns, as they tend to be very long. Third, most block algorithms have very limited databases. GenBank does not store data of this type. Researchers must create their own databases. At the

18

present time, these databases contain only proteins and there is no block intronic information against which to compare a new sequence.

A similar alternative to multiple alignment algorithms is a modeling methodology based on a training set of data. There is overlap between these models and models that aid in mapping, because some models compare multiple genes as well as identify new genes. Models like neural networks (Snyder and Stormo, 1993) and hidden Markov models have been used fairly successfully for comparing multiple sequences of proteins or coding regions (Krogh et al., 1994, Baldi et al., 1994). However, these models require prior knowledge of closely homologous genes to create training sets for the inclusion of future genes; such prior biological knowledge is often not available for intronic regions. In addition, these models are prone to errors even when there are good data (Krogh, 1994). Patterns in the code sometimes cause problems for these algorithms and introns are known to have highly repetitive regions that create unique patterns. Further, these algorithms tend to suppress sequence alignment overlap of more than 15 bp, and introns are generally very variable in length. All of these issues make these models impractical for identifying intronic homologies.

Single Alignment and Homology Between Intronic Regions

By process of elimination, single alignment techniques are the present method of choice when it comes to comparing intronic regions. These methods are probably the most commonly used for any type of DNA analysis. The scenario for their use is fairly consistent. A researcher goes into the laboratory and, with the aid of specific primers, sequences some small portion of the human genome, yielding a clear, readable sequence, whose purpose is unknown. In the absence of an expensive DNA computer system, an

automated homology algorithm like the Basic Local Alignment Search Tool (BLAST),
available free via the INTERNET, will give some idea whether the sequence has already
been identified.

Easy access to DNA analysis tools has helped speed the pace of DNA research, but it
is not without problems. It becomes very important for the researcher to understand the
strengths and weaknesses of the chosen methodology. By design, BLAST does not
attempt to make global (end-to-end) alignments of sequences but instead focuses on
identifying all significant local similarities across a pair of sequences (Boguski, 1992).
This is a deterrent if one has a long intronic sequence and wants to search the database for
another similar long intronic sequence. BLAST will tend to truncate both sequences in
favor of a short region of high homology over two long weak alignments.

Almost all the single alignment algorithms have several features in common (whether
they are commercial or governmental in origin). Most databases rank the possible
alignments using scoring systems which vary by algorithm; many scoring systems created
in the 1980's are now inadequate because they were designed for much simpler
comparisons. Sequences used to be smaller, more uniformly conserved and single
domain protein oriented. A scoring system designed for such sequences is inadequate for
long weakly-conserved regions.

Most government and commercial DNA programs offer local alignment scoring based
on a system that sums the scores for aligned pairs and then scores for gaps. Due to the
lengths of the sequences, the computer processor bogs down quickly. It becomes
necessary to have a machine with a parallel architecture to compute them. To reduce the
computational time, some algorithms (e.g., BLAST) perform the local alignment and

simply forbid gaps. Unfortunately, sensitivity to weak similarities is lost because the focus becomes on smaller, more highly conserved regions. Thus, there is a trade-off: the more complex the algorithm, the higher the price in terms of computation time.

Additionally, even those systems that do score the gaps do not handle introns well. Different types of gaps are important to introns. Short gaps are more indicative of a lack of homology than longer gaps. Although this seems rather counterintuitive, it has to be remembered that DNA folds and bends. This bending and folding tends to occur in intronic regions. As such, a long gap may not indicate a lack of homology because the DNA may bend around and realign with the comparison sequence at a later position. However, when the current algorithms do score the gaps, they weigh every mismatch the same. They do not preferentially treat introns on the basis of their known attributes.

Another problem with these alignment techniques is that it is very difficult to sort out the statistically significant scores from the insignificant ones. This problem is not unique to intronic regions, but it is compounded by the fact that intronic regions, by nature, tend to be more weakly conserved in nature. This means there will be more weak matches with equivalent scores, assuming that one has even captured the biologically important sequences. It is very possible with these algorithms that a sequence of importance is lost among all the irrelevant or chance similarities. The longer a sequence, the greater the score expected by chance (Altschul, et al., 1994). Thus, even though this is the best method currently available (in terms of cost, availability, and ease of use), it has some serious deficiencies when used to investigate homology between two intronic regions. This hole in the available methodologies to characterize and align gene sequences based on their intronic regions needs to be filled.

21

CHAPTER III

DATA COLLECTION

Methodology

Computer Processing and Statistical Applications Software

Data collection and analyses were performed on a MICRON Pentium 90

Powerstation. DNA sequence data were collected using Network Entrez, a TCP/IP-based

client-server version of World Wide Web (WWW) Entrez. Network Entrez is accessed

via the WWW, but its retrieval capability is much quicker than WWW Entrez.

Statistical data analyses used a combination of S-Plus for Windows, Version 3.2

(Release 1) and SAS for Windows, Version 6 (Release 6.10). S-Plus functions were

created for the test for Markov structure, homology, and difference of means. SAS

programs performed the cluster analyses using PROC CLUSTER (SAS, 1988) and

MACRO procedures (SAS, 1989). Appendices D and E contain all SAS and S-Plus

programs used in this thesis.

Criteria for Gene Selection

The data represent an availability sample. All oncogenes, tumor suppressor genes and

non-regulatory genes meeting the following criteria were included: (1) the gene had to

have partial intronic information available for over 50% of its introns, (2) a given intron

had to contain more partial intronic information than merely the intron ends that are used

to match the primer (these can be easily detected because they tend to be 10 to 20 bp long

and fall on either side of each exon), (3) when only a partial intron was sequenced, either

GenBank or a published article had to provide information regarding the total lengths of

the introns in the gene for over 50% of its introns. When a gene had alternative splicing positions, the most biologically established path for that gene was chosen (as provided by GenBank or a published paper).

For this study, it was deemed more important to have valid, more complete information for fewer genes than to investigate corrupted, incomplete information for a large number of genes (Weir, 1993).

In most cases, if a gene included any intronic data, it had partial sequence data available for all introns. In several cases, the author did not know the actual length of one intron, but knew the lengths of over 50% of the remaining introns. Of greater difficulty was the number of genes with no intronic information. From DNA files of 100 oncogenes, 29 tumor suppressor genes, and 40 non-regulatory genes, only 18 oncogenes, 5 tumor suppressor genes and 10 non-regulatory genes met these criteria. More data are expected to be available in the next several years. Hesketh (1995) provides lists of regulatory genes.

Table 3.1 displays the non-regulatory genes (labeled (NR)) and their corresponding acronyms in the data set. Table 3.2 displays the regulatory genes. This table includes the most prevalent types of cancer associated with each gene. Oncogenes are labeled (O) and tumor suppressor genes are labeled (TS).

Table 3.1 Non-regulatory Genes and Their Abbreviations

G6PD(NR)	Glucose-6-phosphate dehydrogenase
PGK(NR)	Phosphoglycerate kinase
ADOL(NR)	Aldolase
GTP(NR)	Phosphoenolpyruvate carboxykinase
PFK(NR)	Phosphofructokinase
TPI(NR)	Triose phosphate isomerase
GCK(NR)	Glucokinase
GAPDH(NR)	Glyceraldehyde-3-phosphate dehydrogenase
SDHB(NR)	Succinate dehydrogenase
HKII(NR)	Hexokinase

(Smith *et al.*, 1983)

Table 3.2: Regulatory Genes and Their Most Common Carcinomas

ABL(O)	Chronic myeloid leukemia
BCR(O)	Chronic myeloid leukemia and Acute lymphoblastic leukemia
FMS(O)	Acute myeloid leukemia
BCL-3(O)	Chronic lymphocytic leukemia
FOS(O)	Osteosarcoma and Acute lymphocytic leukemia
HST-1(O)	Induce blood vessel formation and are synthesized by many tumor cells
INT-2(O)	Amplified in Breast carcinoma and esophageal carcinoma
LCK(O)	Renal cell carcinoma and chronic lymphocytic leukemia
MYC(O)	Small cell lung carcinoma, breast and cervical carcinoma
L-MYC(O)	Breast and colon carcinoma
N-MYC(O)	Neuroblastoma, retinoblastoma and small cell lung carcinoma
MAX(O)	NIH 3T3 fibroblasts, HeLa cells, neuroblastoma-derived cell lines
PIM1(O)	Acute myeloid and lymphoid leukemias
KIT(O)	Small cell lung carcinoma, acute myeloblastic leukemia blast cells
RAFA1	Stomach laryngeal, lung carcinomas and sarcoma
FPS(O)	Lung carcinoma and hematopoietic malignancies
WNT-1(O)	Retinoblastoma
K-RAS2(O)	Lung, colon and pancreas carcinomas
MLH1(TS)	Hereditary nonpolyposis colon cancer
MSH2(TS)	Hereditary nonpolyposis colon cancer
MTS1(TS)	Multiple tumor suppressor
RB(TS)	Retinoblastoma
NF1(TS)	Neurofibrosarcomas and malignant Schwannomas

(Hesketh, 1995)

DNA Sequence Data Collection

Because GenBank is a retrieval system, not a relational database, its use for

exploratory analyses is labor intensive. Moreover, most files in GenBank do not contain

complete intronic information; thus, original papers verified the length and location of the

introns used in this study. Only partial intronic information is often available. For

convenience, an author may have split the DNA code across an intron. Figure 3.1 depicts

a GenBank sample file.

```
LOCUS     HSMYCC     8082 bp  DNA          PRI     27-JUL-1994
DEFINITION  Human c-myc oncogene.
ACCESSION  X00364 J00120 K01908 V00501
NID      g34820
KEYWORDS   myc cellular oncogene; oncogene cellular.
SOURCE    ˙human.
ORGANISM  Homo sapiens
AUTHORS  Gazin,C., Dupont de Dinechin,S., Hampe,A., Masson,J.M., Martin,P.,
        Stehelin,D. and Galibert,F.
 TITLE    Nucleotide sequence of the human c-myc locus: provocative open
        reading frame within the first exon
 JOURNAL  EMBO J. 3 (2), 383-387 (1984)
 MEDLINE  84182501
REFERENCE  2  (bases 3507 to 7559)
 AUTHORS  Colby,W.W., Chen,E.Y., Smith,D.H. and Levinson,A.D.
 TITLE    Identification and nucleotide sequence of a human locus homologous
        to the v-myc oncogene of avian myelocytomatosis virus MC29
 JOURNAL  Nature 301 (5902), 722-725 (1983)
 MEDLINE  83141777
COMMENT    NCBI gi: 34820
FEATURES          Location/Qualifiers
   source     1..8082
            /organism="Homo sapiens"
   CDS        2304..2870
            /note="pot. ORF (aa 1-188);  NCBI gi: 312410"
            /codon_start=1
/translation="MRGSGRLRTPELCCSRPPPPGPGRPWLPSCLEKGRASQRLGGKK
*
   exon       <2304..2881
            /number=1
   intron     2882..4505
            /number=1
   exon       4506..5277
            /number=2
   CDS        join(4521..5277,6654..7216)
            /note="NCBI gi: 34821"
            /codon_start=1
            /product="48K protein"
/translation="MPLNVSFTNRNYDLDYDSVQPYFYCDEEENFYQQQQQSELQ *
   intron     5278..6653
            /number=2
   exon       6654..7657
            /number=3
   polyA_signal  7511..7516
            /note="pot. polyA signal"
   polyA_signal  7652..7657
            /note="pot. polyA signal"
BASE COUNT    1850 a  2115 c  2135 g  1982 t
ORIGIN
      1 agcttgtttg gccgtttag ggtttgttgg aatttttttt tcgtctatgt acttgtgaat
     61 tatttcacgt ttgccattac cggttctcca tagggtgatg ttcattagca gtggtgatag *
```

Figure 3.1: Sample GenBank File from Entrez

*   Amino acid translations were truncated at one line.

** Nucleotide sequence data was truncated at 120 bp.

25

Collection of Intronic Length Data

Tables of intronic lengths were created in MicroSoft Excel, Version 5.0 and transported to S-Plus and SAS as text files.

Data Limitations

This thesis assumes that the information provided by GenBank and published papers is accurate. To the extent possible, I have tried to use the most current GenBank files and verify any obvious discrepancies. I found two obvious mistakes in the code, nucleotides labeled *D* and *M*. Fortunately, the accompanying papers provided the correct nucleotide information and these errors were rectified. In several other cases, the published paper differed from the GenBank sequence length by one or two nucleotides; in all of these instances, sequence numbers were provided by GenBank, since inaccurate computations in the published papers accounted for the discrepancy.

Most sequencing is accomplished by one of two methods -- the Maxam and Gilbert DNA-sequencing procedure or the Sanger DNA-sequencing procedure (Watson, 1992). Today's sequencing process often includes scanning the sequence into a computer file as it is read off the gel. Although there is much progress in this area, sequencing errors still occur for a variety of reasons (Lipshultz, 1994). Compression artifacts can develop when the longer strands of DNA migrate faster than the shorter strands. Temperature variations across the gel (based on different amounts of thickness caused by heat dissipation) can lead to uneven mobility within a band. This will cause the band to broaden. Broad bands tend to overlap and are more difficult to read. Fragments longer than 400-500 bp are often more difficult to read.

Manual trace editing can find some errors, but this takes time and costs money. Some labs have now incorporated an automated trace editing (tedding) tool to replace the manual process. This can cheaply eliminate errors, but very few labs have this capability (Lipshultz, 1994).

Lipshultz estimates that an automated sequencer has an error rate of 4% relative to the consensus sequence. Roberts and Posfai developed a program called DETECT (Roberts, 1991). Using DETECT, they found 156 clear errors in 1.3 million bases due to frame-shift. A frame-shift error occurs when there is an addition or substitution of a nucleotide during DNA replication such that the normal reading frame based on nuclotide triplets is shifted. This results in translation of the wrong sequence of amino acids after the point in the code where the change occurs (King and Stansfield, 1990). They estimate that 1% of GenBank has frame-shift errors and 5% has substitution errors.

The frame-shift problem applies to the accurate translation of the nucleotide code into the proper amino acids based on grouping of three nucleotides. If one of the three nucleotides is deleted, or an extraneous nucleotide is added, a frame-shift occurs. Frame shift-errors are not as big a problem for intronic nucleotide sequence data, although the code still has an incorrect insertion or deletion. But the 4-5% substitution rate error means that one cannot claim that DNA sequence data entered into GenBank is flawless.

There is also a degree of error due to laboratory estimation of intronic lengths for regions that have not been sequenced. The lengths provided are generally rounded; e.g., a sequence of length 1294 may be recorded as 1300. As more intronic regions are sequenced, such rounding may occur less frequently.

CHAPTER IV

INITIAL CLUSTER ANALYSES

Methodology

In this chapter, I investigate whether different categories of genes (i.e., regulatory and non-regulatory genes) have intronic features so closely related that they will cluster. A cluster is a set of objects (e.g., genes) that resemble each other more than they resemble other objects not in the cluster. New objects joining a cluster are expected to have properties similar to those already in the cluster.

The features on which I base gene clustering in this thesis are: the number of introns in a gene, the mean of the log lengths for introns in a gene, and the standard deviations of the log lengths for introns in a gene.

Classes of Clustering Methods

Clustering methods are organized into two classes: partitioning methods and hierarchical methods. The choice of clustering method depends upon the underlying purpose behind the research.

In partitioning, the data are classified into K groups, where K is determined in advance (Kaufman, 1990). A problem with partitioning is that the algorithm provides the number of clusters requested by the researcher, not necessarily the optimal number of clusters. Additionally, partitioning does not yield insight into the hierarchical relationships between objects within a given partition. Some common types of partition methods include k-means clustering and fuzzy algorithms.

Hierarchical clustering, also known as tree structuring (Hartigan, 1975), is a series of successive fusions of the attributes into groups (Everitt, 1980). Hierarchical algorithms consider all possible partitions from $K = 1$ to $K = n$ and allow visualization of the intra-group relationships within each partition. Since the biological intra-group relationships between genes are important to this thesis, hierarchical methods are used.

There are two kinds of hierarchical techniques: agglomerative and divisive (Kaufman, 1980). These methods construct the hierarchy in opposite directions. Divisive methods start at step zero with all the objects in one group. In each subsequent step, a cluster is split until there are n clusters at the end of the process. Agglomerative methods start at step zero with n clusters. At each step two clusters are merged until there is one cluster at the end of the process. The agglomerative technique is used in this thesis.

Clustering Algorithm Input Structure

Clustering algorithms use one of two input structures, a two-mode structure and a one-mode structure (Kaufman, 1990).

Two-mode Structure. The two-mode structure is the most common. The first step is to create an $n \times p$ coordinate data matrix, where n is the number of objects and p is the number of features. The second step is to calculate a quantitative distance measure between each pair of objects and create an $n \times n$ distance matrix. The most notable quantitative distance measure uses the Euclidean distance:

$$d_{ij} = (\sum_{k=1}^{p} (x_{ik} - x_{jk})^2)^{1/2}$$

where x_{ik} is the value of the kth feature for the ith object. Euclidean distances will be used in this chapter. Other common measures are squared Euclidean distance, Manhattan distance (or city block) and Minkowski distance (Everitt, 1980).

One-mode Structure. This structure is a direct collection of distances based on all pairs of objects. These distances combine to create an n x n matrix. One mode structures are sometimes created in place of the common distance measures discussed above.

A common misconception about cluster analysis is that it always must begin with an *object* x *feature* matrix (i.e., an n x p matrix). Many times the one-mode structure is more applicable. In this thesis, two-mode structures are used in this chapter and a one-mode structure is created in chapter 6.

Hierarchical Clustering Methods

Hierarchical clustering methods differ primarily in how distance is determined from one object to another object or cluster. However, the basic steps in an agglomerative hierarchical clustering are common to all methods (Johnson and Wichern, 1988). These are:

(1) Start with N clusters where each cluster contains a single object

(2) Create an N x N symmetric matrix of distances, $D = \{d_{ik}\}$

(3) Find the nearest pair of clusters (U and V) in the distance matrix

(4) Designate the distance between clusters U and V as d_{UV}

(5) Merge clusters U and V to form *(UV)*

(6) Update the distance matrix by deleting the rows and columns associated with U and V

(7) Add a row and column providing the distances between cluster *(UV)* and remaining clusters

(8) Repeat steps 3 - 7 *N-1* times; record the order of the mergers.

Choice of Clustering Method

One of the hardest decisions in a cluster analysis is the choice of an appropriate clustering method. No clustering method is optimal in all situations (Milligan, 1995), with respect to its ability to find true clusters within the data. The researcher is faced with numerous methods, all with inherent strengths and weaknesses. These strengths and weaknesses have been identified by a compilation of simulation studies that have been used to determine situational best methods based on different data scenarios (Kuiper and Fisher, 1975, Milligan, 1995). When there is near perfect information about a data set, it may be possible to choose the most appropriate method. But if the underlying structure of the data is not known (as often occurs in exploratory biological analyses), choice of a best method becomes extremely difficult. As a solution, it is a recommended that one use numerous methods and look for a consensus among results (Everitt, 1980, SAS, 1988, Johnson and Wichern, 1988, Milligan, 1995). Table 4.1 provides a description of the nine clustering methods offered by SAS that will be employed in this thesis. If the outcomes from the various methods are roughly consistent (i.e., there is a consensus among methods concerning cluster members), a case for natural groupings can be made. This approach (i.e., drawing conclusions from the nine clustering methods) is followed in this thesis, with one area of modification based on results from Milligan's work.

Milligan (1980, 1985, 1995) has done a substantial amount of research in the area of choosing optimal clustering methodologies. He has examined a number of methods

31

Table 4.1: Description of Clustering Methods Used in This Thesis

Clustering Method	How Distance is Determined	Strengths	Weaknesses
Single Linkage (Nearest Neighbor Method)	Groups of objects merge according to distance between their nearest members (Everitt, 1980)	- One of the few methods that can describe non-ellipsoid clusters (Johnson and Wichern, 1988) - Can detect elongated and irregular clusters (SAS, 1988)	- Consistently performs poorly, even when the data are known to be error free (Milligan, 1981) - Sensitive to almost any type of error added to original data (Milligan, 1995) - Tendency to chop-off tails of distribution before main clusters separate (SAS, 1988) - Unable to discern poorly separated clusters (Kuiper and Fisher, 1975) - Has a tendency to "chain", i.e., to combine objects in a chain-like structure into a single cluster; can be misleading if opposite ends of chain are dissimilar (Kuiper and Fisher, 1975)
Complete Linkage (Furthest Neighbor Method)	Groups of objects merge according to the distance between their farthest members (Everitt, 1980)	- For compact clusters of nearly equal size, complete linkage and Ward's are best (Kuiper and Fisher, 1975) - Somewhat robust against the number of clusters, and the separation of clusters (Kuiper and Fisher, 1975)	- Strongly biased toward producing clusters with roughly equal diameters (Milligan, 1980) - Can be severely distorted by moderate outliers (Milligan, 1980) - Not robust to errors in measurement of features (Fisher and Kuiper, 1975)
Average Linkage (Unweighted Arithmetic Average Clustering (UPGMA))	Distance between two clusters is the average distance between all pairs of objects where one member of a pair belongs to each cluster (SAS, 1988)	- Preferable when there is wide disparity in number of points in clusters (Kuiper and Fisher, 1975) - Is competitive in terms of cluster recovery (but isn't as quite as consistent as Ward's and beta-flexible) (Milligan, 1995)	- Tendency to join clusters with small variances (SAS, 1988) - Slightly biased toward producing clusters with the same variance (SAS, 1988).
Ward's Minimum-Variance Method	At each clustering generation, the within-cluster sum of squares is minimized over all the partitions created by the two clusters that merged during the previous generation (SAS, 1988).	- Advantageous over other methods in different simulation scenarios: error induced on data points, different similarity measures, introduction of outliers, different population distributions, different cluster sizes, number of variables defined is varied (SAS, 1988 and Milligan, 1995) - Somewhat robust against number of clusters and separation of clusters (Kuiper and Fisher, 1975) - For compact clusters of nearly equal size, complete linkage and Ward's are best (Kuiper and Fisher, 1975)	- Tends to be very sensitive to outliers (Milligan, 1980) - Tends to merge clusters with small number of observations (SAS, 1988) - May be strongly biased toward producing clusters with same number of observations (SAS, 1988) - Not always best for unequal sample sizes, even when data are bivariate normal (Kuiper and Fisher, 1975) - Assumes distribution of features on each object is multivariate normal, equal covariance matrices, and equal sampling probabilities; difficult assumptions to achieve (SAS, 1988)

Table 4.1 (Continued): Description of Clustering Methods Used in This Thesis

Clustering Method	How Distance is Determined	Strengths	Weaknesses
Centroid Method (Unweighted Centroid Clustering (UPGMC))	Distance between two clusters is defined as the distance between their centroids or means (SAS, 1988)	- More robust to outliers than most other hierarchical methods (Milligan, 1980) - Performance improves if assumption of equal group size is made; forces position of new group to be between two previous groups (Everitt, 1980) - Preferable when there is wide disparity in number of points in clusters (Kuiper and Fisher, 1975)	- Does not perform as well in simulation studies as Ward's or average linkage (Milligan, 1980). - Disadvantageous when merging two clusters of very different sizes because centroid of new group may be closer to larger group; features of the smaller group may be lost (Everitt, 1980)
Beta-Flexible Method	Derived from a general model that approximates the other hierarchical clustering methods by modifying (flexing) the parameter β.	- Based on simulation studies (refer to Ward's method) this method is advantageous to all others except Ward's (Milligan, 1995) - β is adjustable, especially for data with a lot of outliers (SAS, 1988)	- Theoretical origin of this model is ill-explained in literature - Rationale for selection of values for β are ill-explained
McQuitty's Method (Weighted Arithmetic Average Clustering (WPGMA) Method)	Similar to average linkage, except it is a weighted method. It averages the similarities between all the members of the two clusters about to fuse, giving the same weight to each branch.	- Method is stronger than average linkage when quality of the sampling scheme is questionable (Legendre, 1995)	- Method is weaker than average linkage when quality of sampling scheme accurately represents nature (Legendre, 1995)
Equal Variance Maximum Likelihood (EML) Method	Joins clusters to maximize the likelihood (based on multivariate normal) at every level of the hierarchy.	- EML removes the bias toward equal size cluster that is evident with Ward's method (SAS, 1988) - Created to handle disjoint clusters (SAS, 1988)	- Assumes distribution of features on each object is multivariate normal, equal covariance matrices, and unequal sampling probabilities; difficult assumptions to achieve (SAS, 1988) - Is sometimes slightly biased toward unequal-sized clusters (SAS, 1988) - Can only be used with coordinate data - Not well studied; SAS unique product (SAS, 1988)
Median Method	Distance between previously combined clusters (UV) and new cluster W lies along the median of a triangle created by the three groups.	- Method appears to be least used, despite the fact that it performs fairly well when the centers of the clusters are uniquely defined (Kaufman, 1990).	- Centers of clusters cannot always be uniquely defined; result is method fails to get same attention as other methods (Kaufman, 1990)

(mostly agglomerative hierarchical methods) using Monte Carlo simulations. His validation studies are designed to examine the effects of various factors on cluster groupings including: inducing error on the data points, testing different similarity measures, introducing outliers in the data set, assuming different population distributions, evaluating different cluster sizes, and varying the number of features included in the clustering process (Milligan, 1995). Based on his validation testing, he recommends Ward's minimum-variance method, and the Lance and Williams beta-flexible approach. Average linkage method also fared well in validation testing, although its performance is not as consistent as Ward's and beta-flexible. Single linkage consistently performed poorly.

In this thesis, I follow the published advice of Everitt (1980), Johnson and Wichern (1988) and Milligan (1995) by performing cluster analyses for all methods mentioned above. I use the results from all methods to determine the proper number of clusters, but preferentially emphasize the results based on the three methods recommended by Milligan (1995).

Determining the Number of Clusters

A common misconception of clustering is that there is an optimal number of clusters. Most clustering methods were not designed to determine the number of clusters in the data. There are methods for investigating this issue, but there is no guarantee of a consistency between methods and definitive resolution of the correct number.

It is tempting to determine the optimal number of clusters through a hypothesis testing approach to assess the statistical significance of a cluster structure (Milligan and Cooper, 1985), e.g., via analysis of variance (ANOVA) or multiple analysis of variance

34

(MANOVA). Such an approach is inappropriate because the results will almost always conclude significant mean differences, even when the data are random noise and no true clusters are present. Cluster analysis inherently disobeys the underlying assumptions behind ANOVA and MANOVA, namely that the data are independently sampled from a multivariate normal distribution and that the objects are allocated to the clusters randomly. Hierarchical clustering meets neither of these assumptions. Additionally, cluster analysis does not have an underlying assumption of normality (SAS, 1988, and Milligan, 1995). For these reasons, standard hypothesis testing methods are inappropriate.

There have been numerous attempts in the last thirty years to address the optimal number of clusters issue. These approaches are called stopping rules, indicating the place where one is to stop in the clustering hierarchy. A comparative study of thirty attempts was made by Milligan and Cooper (1985). Milligan and Cooper simulated data that had an error free structure and in which distinct clustering was present. They analyzed various data sets, examining the ability of the stopping rules to detect 2, 3, 4, or 5 clusters. Based on the results of this simulation study, Milligan recommends that only the top five stopping rules be used to determine the proper number of clusters. Ideally, at least two stopping rules should be used simultaneously. With concurrence between methods, one has the best chance of determining the correct number of clusters. SAS incorporates the methods ranked first and second from Milligan and Cooper's recommended list, namely the pseudo F and the pseudo t^2.

Pseudo F.   The formula for the pseudo F at a given level in the hierarchy is

$$\text{pseudo } F = ((T - P_G)/(G - 1))/(P_G/(n-G))$$

where

$$P_G = \sum_{U=1}^{G}\left[\sum_{i \in C_U}\left[\sum_{k=1}^{p}(x_{ik} - \bar{x}_{.k})^2\right]\right] \quad \text{(where } C_U \text{ is the } U^{th} \text{ cluster)}$$

$$T = \sum_{i=1}^{n}\left[\sum_{k=1}^{p}(x_{ik} - \bar{x}_{.k})^2\right]$$

G = the number of clusters at the Gth level in the hierarchy (SAS, 1988).

An equivalent way of representing the pseudo F is to state that it is the ratio of trace (sum of the objects on the main diagonal) of the between-cluster sum of squares matrix divided by $(G-1)$ and the trace of the within-cluster sum of squares divided by $(n-G)$. F is maximized for a given level in the hierarchy when the within-cluster sum of squares decreases. If the data have been collected independently from a multivariate normal distribution and randomly allocated to groups, the pseudo F is distributed as an F distribution with $p(G-1)$ and $p(n-G)$ degrees of freedom. But since hierarchical cluster analysis does not require a normal distribution, nor does it randomly allocate objects to groups, the pseudo F is not distributed as an F distribution. Nonetheless, it is still a useful indicator of the correct number of clusters (SAS, 1988). The pseudo F does not rely on a critical value to determine the optimal number of clusters but instead relies on a local maximization of the pseudo F at each level of the cluster hierarchy. The chosen pseudo F for the optimal number of clusters will be higher than the pseudo F's for the surrounding clusters (i.e., the within sum of squares is minimized).

Pseudo t^2. The pseudo t^2 formula for joining C_U and C_V is

$$\text{pseudo } t^2 = B_{UV} / ((W_U + W_V) / (N_U + N_V - 2))$$

where

$$B_{UV} = W_{UV} - W_U - W_V \quad \text{if } C_{UV} = C_U \cup C_V$$

$$W_U = \sum_{i \in C_U} \left[\sum_{k=1}^{p} (x_{ik} - \bar{x}_{.k})^2 \right]$$

N_U = number of observations in C_U

This formula creates a ratio. The denominator is the sum of squares for error when the data are partitioned into two clusters. The numerator is the squared error of the new cluster minus the squared errors of the two separate clusters. The pseudo t^2 is optimal when the ratio is minimized (Milligan and Cooper, 1985). The pseudo t^2 is distributed as an F random variable with p and $p(N_U + N_V - 2)$ degrees of freedom if the data are independently sampled from a multivariate normal distribution and the objects are assigned randomly to clusters (SAS, 1988). The pseudo t^2 is computed only from information in the latest cluster merger. This means the effective sample size is reduced compared to the overall sample size, so the statistic may suffer from a lack of power to detect significant differences. However, Milligan and Cooper's (1985) simulation studies determined that the pseudo t^2 consistently performs well and offers an effective technique for situations where the sample size is not large.

Selection of Variables

One of the beneficial features of cluster analysis is the ability to combine different types of variables from different types of data into the analysis. Variables can be counts,

ratios, intervals, ordinal or categorical (Hartigan, 1975). Cluster analysis can also handle missing values. In cluster analysis, one should not arbitrarily bring many variables into the process because the addition of even one or two irrelevant variables can dramatically inhibit cluster recovery (Milligan, 1995).

Decision to Standardize

If sizable differences in means and variances exist between variables, standardization is appropriate. If they do not exist, standardization will not add much to the process (Milligan and Cooper, 1985). In this thesis, standardization was incorporated based on a zero mean and a standard deviation of one (SAS, 1988).

Criticisms Levied at Cluster Analysis

Two complaints about the clustering process are: (1) cluster analysis is not based on sound probability models, and (2) results are poorly evaluated and unstable.

Hartigan (1975) counters the first complaint by stating that most statistical analyses compare predetermined groups: the objects are assigned to the groups because the researcher assigned them, not necessarily because the objects optimally belong in those groups. This approach works fine when the underlying structure of the data is known. But, when the underlying structure is unknown, the appropriate use of these methods becomes questionable. An advantage of cluster analysis is that it offers a methodology that does not expect predetermined groupings, providing the ability to observe what may be closer to the real state of the data in nature. Since an underlying structure is usually not implied, it remains credible under differing data structures.

Although individual clustering methods may appear to be unstable and poorly evaluated, the stability of cluster analysis can be greatly enhanced by looking for

38

consensus among several hierarchical methods. This consensus supports the assurance of stability that a single hierarchical method cannot. Additionally, there have been numerous simulation studies to evaluate hierarchical models under diverse scenarios; these provide additional information that can improve the stability of the analysis.

Application of Cluster Analysis to Intronic Features Data

Cluster analyses were performed for all combinations of the three biological features. Of the seven possible combinations, two provided biologically significant results that I will further discuss in this thesis: clustering based on number of introns in a gene, and clustering based on both number of introns and mean of the intronic log lengths in a gene. The standard deviation of intronic log lengths in a gene did not result in biologically significant clusters, either when analyzed separately or when combined with other features.

The data are depicted in table 4.2. The number of introns for the 33 genes in the data set varies dramatically. The range is 1 to 56 introns per gene. Plots of the data (truncated at 16,000 bp for better visualization) by regulatory group are provided in figures A.1-A.4 of appendix A. Because the intronic lengths have such a wide range, the natural logarithm was used as a transformation function where

$$Z_{ij} = \ln (X_{ij}).$$

Transformations of this type are considered acceptable in cluster analysis (Romesburg, 1984). Non-monotonic transformations are not acceptable. They can change the number of population clusters and should be approached with caution (SAS, 1988). Simultaneous use of transformations and standardization is permitted (Romesburg, 1984). Plots of the transformed data by regulatory group are depicted in tables A.5-A.8 of appendix A. The

Table 4.2: Total Lengths For Introns in Each Gene (Per Data in Published Papers)

Gene	# of Introns In Gene	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
ABL(O)	10	200000	563	7666	9093	646	1830	2297	1500	342	3306						
BCR(O)	22	71559	6969	300	6867	3016	1488	500	10202	934	1957	818	1344	717	2127	2385	14268
FMS(O)	21	26000	5349	430	1803	676	3780	2370	153	118	1542	6355	127	517	999	2107	119
BCL-3(O)	8	2300	5000	676	105	230	522	310	606								
FOS(O)	3	753	431	114													
HST-1(O)	2	617	538														
INT-2(O)	2	2289	5648														
LCK(O)	11	304	172	236	143	241	261	117	3400	89	84	5800					
MYC(O)	2	1624	1376														
L-MYC(O)	2	364	2971														
N-MYC(O)	2	894	2636														
MAX(O)	1	483															
PIM1(O)	5	113	101	93	1504	761											
KIT(O)	20	NA	2394	960	4708	3973	2557	NA	2014	1019	91	280	83	1349	2163	476	1227
RAFA1(O)	15	1652	162	1458	85	114	1290	103	74	98	1281	91	505	236	853	308	
FPS(O)	18	492	152	1373	129	1931	73	203	129	121	495	361	381	594	274	96	239
WNT-1(O)	3	713	702	462													
K-RAS2(O)	4	5746	12941	1651	15581												
MLH1(TS)	18	2900	4000	3300	2500	1600	2900	154	2300	2900	2700	5500	2300	12600	1900	5100	800
MSH2(TS)	15	4400	1600	1900	1800	1900	13000	9300	17000	3500	4000	3800	1000	1700	1900	1800	
MTS1(TS)	3	231	658	721													
RB(TS)	26	3430	33550	2048	2648	1829	12022	2313	2232	2338	1581	3806	3624	2367	403	80	1550
NF1(TS)	56	80000	3100	NA	NA	2000	220	800	NA	400	NA	4000	NA	4000	2500	540	NA
G6PD(NR)	12	625	9856	95	549	671	177	365	447	139	104	105	97				
PGK(NR)	10	5751	3860	261	3255	700	4727	278	1457	273	359						
ADOL(NR)	8	4811	809	1231	836	841	778	411	2937								
GTP(NR)	8	435	440	128	350	440	89	424	214								
PFK(NR)	21	9000	1500	1100	1800	530	150	2650	92	1900	750	250	400	900	800	800	900
TPI(NR)	6	1165	111	74	297	272	128										
GCK(NR)	9	9012	4600	1400	1900	1100	2000	1200	700	900							
GAPDH(NR)	8	240	1634	90	129	90	92	193	104								
SDHB(NR)	7	8800	11700	4900	750	3500	1300	3700									
HKII(NR)	17	15800	6400	4800	800	400	300	2700	1200	1400	1300	200	3600	600	96	242	133

Table 4.2 (Continued): Total Lengths For Introns in Each Gene (Per Data in Published Papers)

Gene												Introns								
	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36
ABL(O)																				
BCR(O)	840	1263	1050	946	478	718														
FMS(O)	945	81	690	806	97															
BCL-3(O)																				
FOS(O)																				
HST-1(O)																				
INT-2(O)																				
LCK(O)																				
MYC(O)																				
L-MYC(O)																				
N-MYC(O)																				
MAX(O)																				
PIM1(O)																				
KIT(O)	3543	111	350	1037																
RAFA1(O)	124																			
FPS(O)		1357																		
WNT-1(O)																				
K-RAS2(O)																				
MLH1(TS)	300	1500																		
MSH2(TS)																				
MTS1(TS)																				
RB(TS)	70488	3542	3239	4582	761	93	8034	3394	511	2304										
NF1(TS)	1200	490	231	1300	369	285	461	1200	550	120	2200	145	NA	4000	530	1250	1270	NA	47500	1300
G6PD(NR)																				
PGK(NR)																				
ADOL(NR)																				
GTP(NR)																				
PFK(NR)	200	300	280	95	443															
TPI(NR)																				
GLUC(NR)																				
GAPDH(NR)																				
SDHB(NR)																				
HKII(NR)	150																			

41

Table 4.2 (Continued): Total Lengths For Introns in Each Gene (Per Data in Published Papers)

Gene	Introns																			
	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56
ABL(O)																				
BCR(O)																				
FMS(O)																				
BCL-3(O)																				
FOS(O)																				
HST-1(O)																				
INT-2(O)																				
LCK(O)																				
MYC(O)																				
L-MYC(O)																				
N-MYC(O)																				
MAX(O)																				
PIM1(O)																				
KIT(O)																				
RAFA1(O)																				
FPS(O)																				
WNT-1(O)																				
K-RAS2(O)																				
MLH1(TS)																				
MSH2(TS)																				
MTS1(TS)																				
RB(TS)																				
NF1(TS)	2700	4300	1550	150	400	240	150	570	1700	2400	6000	930	2000	4000	350	180	1100	350	1400	13296
G6PD(NR)																				
PGK(NR)																				
ADOL(NR)																				
GTP(NR)																				
PFK(NR)																				
TPI(NR)																				
GLUC(NR)																				
GAPDH(NR)																				
SDHB(NR)																				
HKII(NR)																				

variation of the intronic lengths on the transformed scale is greatly reduced, allowing the transformed attributes to contribute more equally to the overall similarity between objects. The means and standard deviations of the intronic log lengths for all 33 genes are shown in table A.1 of appendix A.

The data are placed in an $n \times p$ features matrix where n is the number of objects (i.e., the 33 genes) and p is the number of features (i.e., the three variables of biological interest). All variable combinations are analyzed by the same clustering process. A SAS MACRO was written to perform the nine methods outlined in table 4.1; the SAS code provided in appendix D. Results from the pseudo F and pseudo t^2 for all methods are compared to determine a consensus (when possible) as to the optimal number of clusters. Clustering results for Milligan's three preferred methods are displayed and consensus concerning cluster members is evaluated based on these three methods.

Clustering Based on Number of Introns in a Gene

The clustering process based on the number of introns in a gene is evaluated for the nine clustering methods offered by SAS. As an example, table 4.3 depicts an edited SAS output for the average linkage method based on the number of introns in a gene. On the left side of the table, the order in which the genes enter the clusters is presented. Focusing on the pseudo F column, F is locally maximized (i.e., the within-cluster sum of squares is minimized) as compared to the surrounding cluster levels when the number of clusters is three and thirteen.

The pseudo t^2 is locally minimized when the within-cluster sum of squares is minimized. One should look for dramatic drops when evaluating the pseudo t^2. In this

case, the t^2 drops when the number of clusters is 3, 5, 6 and 10. The most dramatic drop occurs when the number of clusters is 3.

There are a large number of ties in these data, but this is not an uncommon occurrence with discrete data like number of introns in a gene. A tie occurs when two or more pairs have the same minimum distance. PROC CLUSTER breaks ties based on the numbered order of the variables in the data set. Each tied pair has a smaller ID number and a larger ID number. The pair with the minimum larger ID number is chosen. If a tie still results, the pair with the smallest ID number is fused first (SAS, 1988).

Table 4.3: SAS Sample Output (Edited) to Demonstrate Use of the Pseudo F and Pseudo t^2

Average Linkage Cluster Analysis

Number of Clusters	-----Clusters Joined-----		Frequency of New Cluster	Pseudo F	Pseudo t**2	Tie
32	HST-1(O)	INT-2(O)	2	.	.	T
31	CL32	MYC(O)	3	.	.	T
30	CL31	L-MYC(O)	4	.	.	T
29	CL30	N-MYC(O)	5	.	.	T
28	FOS(O)	WNT-1(O)	2	.	.	T
27	FPS(O)	MLH1(TS)	2	.	.	T
26	RAFA1(O)	MSH2(TS)	2	.	.	T
25	CL28	MTS1(TS)	3	.	.	T
24	ABL(O)	PGK(NR)	2	.	.	T
23	BCL-3(O)	ADOL(NR)	2	.	.	T
22	CL23	GTP(NR)	3	.	.	T
21	FMS(O)	PFK(NR)	2	.	.	T
20	CL22	GAPDH(NR)	4	.	.	
19	BCR(O)	CL21	3	4292.13	.	T
18	CL25	CL29	8	1276.52	.	T
17	CL24	LCK(O)	3	1145.90	.	T
16	PIM1(O)	K-RAS2(O)	2	1123.43	.	T
15	CL20	GCK(NR)	5	1048.10	.	T
14	TPI(NR)	SDHB(NR)	2	1072.33	.	T
13	CL27	HKII(NR)	3	**1078.99**	.	
12	CL19	KIT(O)	4	1000.43	4.00	
11	CL18	MAX(O)	9	929.47	6.27	
10	CL17	G6PD(NR)	4	870.39	**6.25**	
9	CL15	CL14	7	737.83	15.88	
8	CL11	CL16	11	558.26	18.84	
7	CL26	CL13	5	495.14	38.40	
6	CL10	CL9	11	353.38	**25.81**	
5	CL12	CL7	9	254.75	**26.89**	
4	CL6	CL8	22	105.60	95.14	
3	CL5	RB(TS)	10	**138.98**	**7.36**	
2	CL4	CL3	32	39.20	105.99	
1	CL2	NF1(TS)	33	.	39.20	

Ties are generally not desirable because the level in the cluster hierarchy where the tie occurs and sometimes subsequent levels are not unique. But, the magnitude of their impact depends upon where in the clustering hierarchy they occur; ties early in the cluster history generally have little effect on the later stages, ties in the middle may signal the need for further investigation, and ties late in the data set are very important. They can signal indeterminacies within the data (SAS, 1988). In the case of these data, the ties occur early enough in the clustering process that they do not pose a problem for the analysis.

The pseudo F and pseudo t^2 results from the other eight clustering methods are examined and reported in table 4.4. The modal number of clusters, 3, is taken as optimal in this analysis.

Table 4.4: Optimal Number of Clusters Based on Number of Introns in 33 Genes

Method	Pseudo F	Pseudo t^2
Ward's	None	4 or 9
Beta-Flexible	None	4,7,9,12
Average	3,13	3,5,6,10
Centroid	3,13	3,5,6,10
Complete	3,13	3,8,11
EML	None	4,6,9,12
McQuitty's	3,13	3,7,11
Median	3,13	3,7,11
Single	2,4,8,10,12	2,4,8,10,12

A disadvantage of SAS is its inability to graphically display the clustering results. PROC TREE is difficult to format and interpret; it does not produce the dendrograms (two-dimensional inverted tree structures that illustrate fusions at each successive level in

45

the hierarchy) commonly associated with cluster analysis. To overcome this limitation,

PowerPoint Version 4.0 was used to depict the clustering results.

Clusterings based on Ward's, beta-flexible and average linkage are shown in figures
4.1, 4.2, and 4.3 respectively. Each method places the same genes in the three clusters.
Thus, there is 100 percent consensus as to cluster members based on the three methods
for the 3 cluster model.
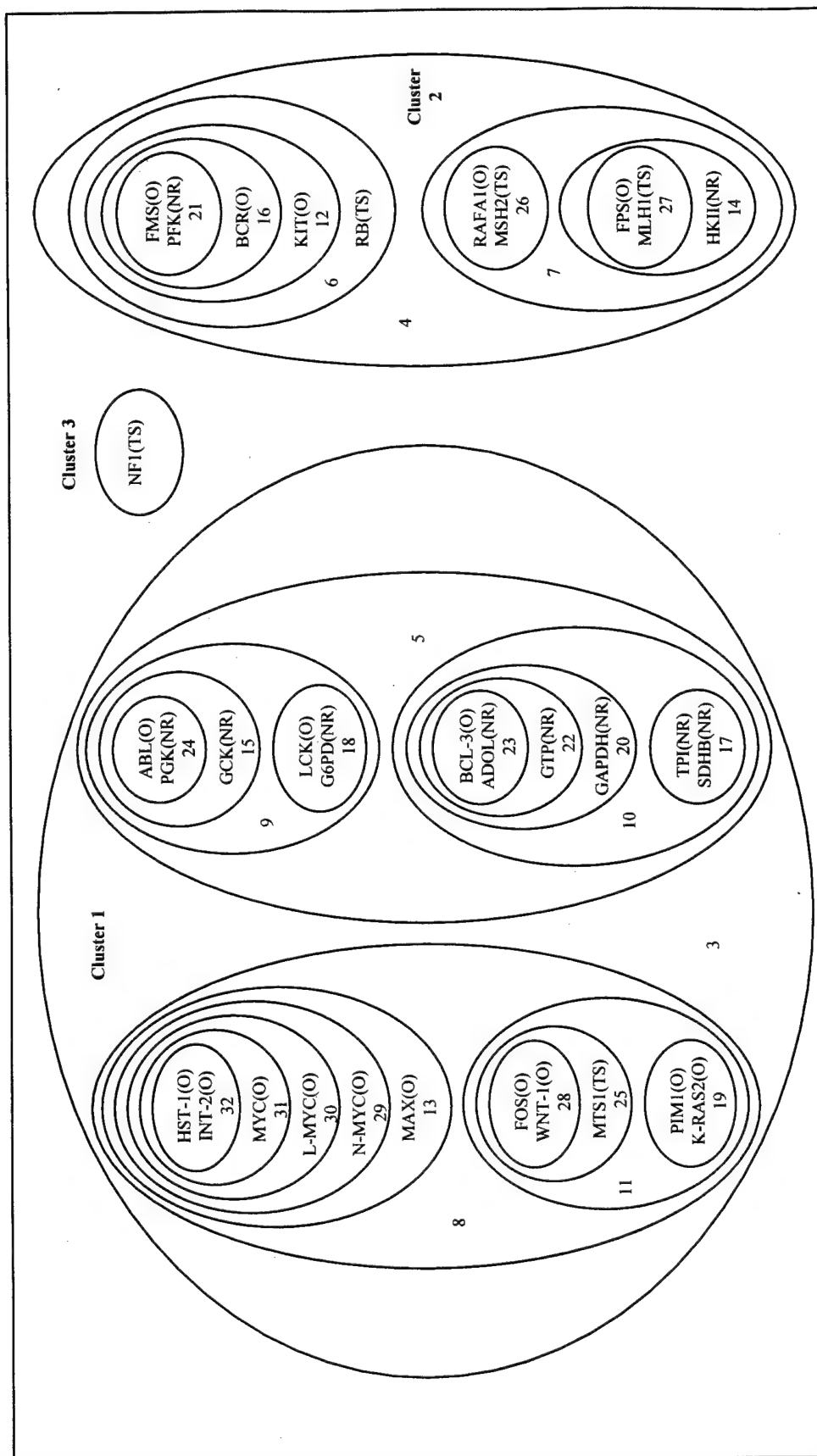
Figure 4.1: Clusters Based on Number of Introns Using Ward's Minimum Variance Cluster Analysis
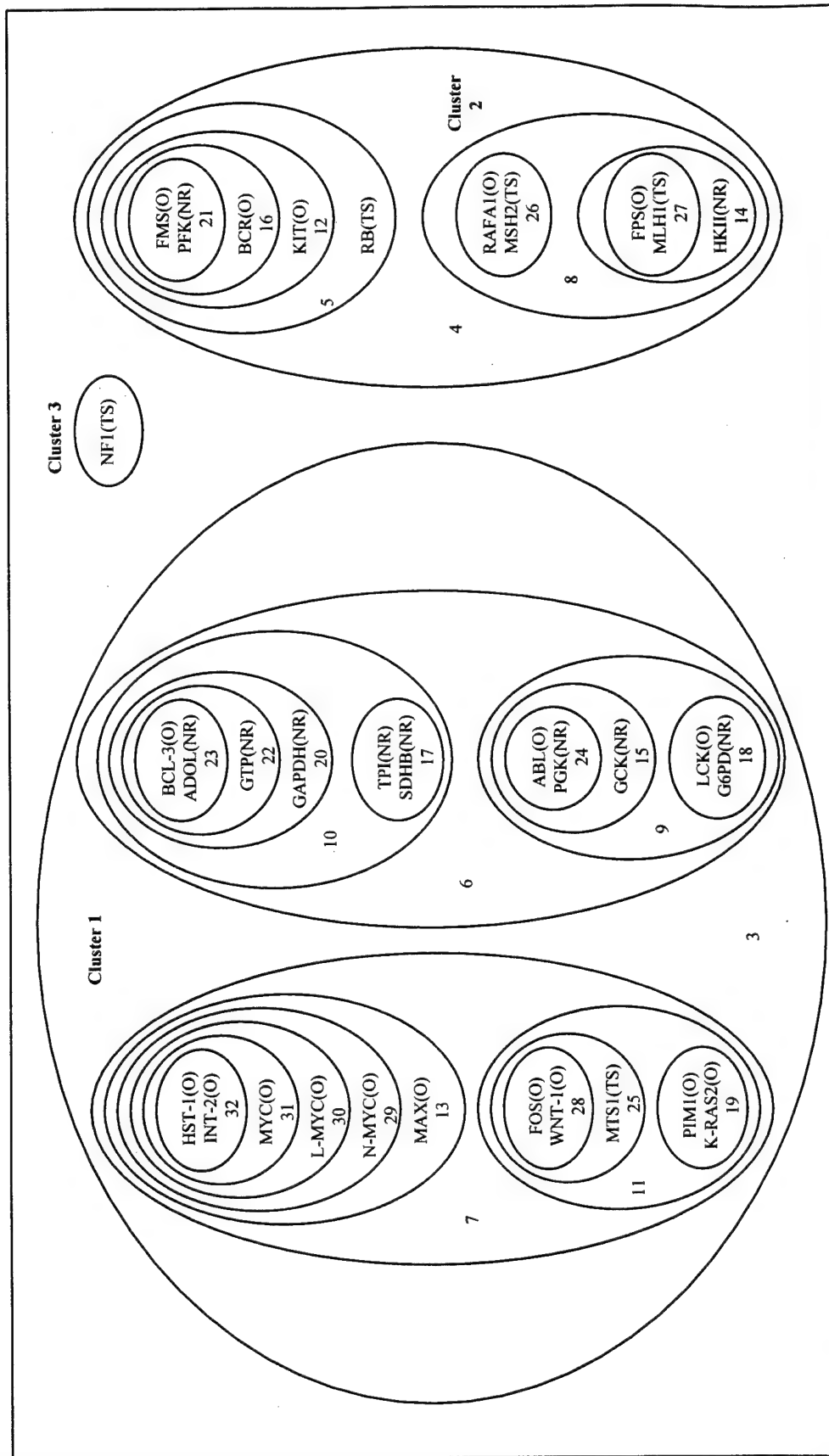
Figure 4.2: Clusters Based on Number of Introns Using Beta-Flexible Cluster Analysis
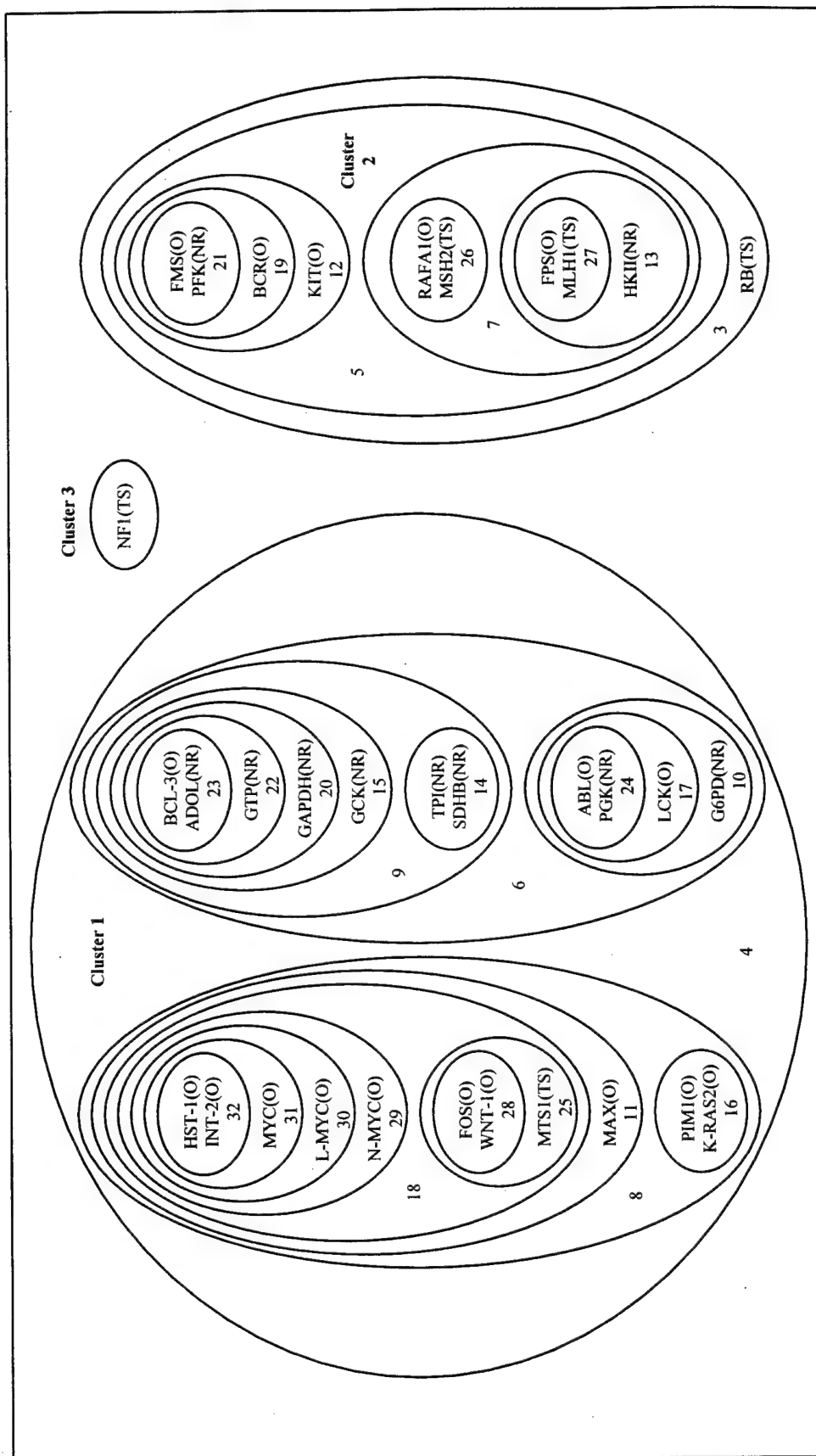
Figure 4.3: Clusters Based on Number of Introns Using Average Linkage Cluster Analysis

Interpretation of Clustering Based on Number of Introns in a Gene

Observing the three clusters created by each method depicted in figures 4.1 - 4.3, cluster 1 contains genes with 1 to 12 introns. Cluster 2 contains genes with 15 to 26 introns. Cluster 3 contains one gene with 56 intronic regions. Within cluster 1, there are two well defined partitions. The first partition consists of almost all oncogenes (tumor suppressor *MTS1* is the exception). Genes in this partition contain 1 to 5 introns. Eight of the ten non-regulatory genes fall into the second partition; these genes contain from 6 to 12 introns.

Within cluster 2, there is more of a gene mixture with two distinct partitions. The number of introns in genes from the first partition range from 15 to 18. This group of genes contains two tumor suppressor genes, two oncogenes and one non-regulatory gene. The genes in the second partition contain primarily 21 to 22 introns. The retinoblastoma *(RB)* gene has a greater number of intronic regions (26) than other cluster members in this partition. When observing the average linkage method, *RB* is slightly removed from the two partitions within cluster 2; the partitions merge before *RB* joins the cluster. This degree of differentiation is not apparent in the other two clustering methods. However, *RB* is not different enough from other members of cluster 2 that it instead joins cluster 3. This signifies that tumor suppressor peripheral neurofibromatosis *(NF1)* is different enough from the other genes (i.e., it has 56 introns) that it forms its own cluster. In fact, all other genes in the data set fuse into a cluster before *NF1* joins the cluster at the last level in the hierarchy.

A partition that includes 8 of the 10 non-regulatory genes appears non-random. This suggests that number of intronic regions may be somewhat different between non-

regulatory and regulatory genes. A partition where 10 out of 11 genes are oncogenes also appears non-random.

Focusing on relationships within this oncogene partition, it is interesting to note that the *HST-1* and *INT-2* genes cluster together. Both are members of the fibroblast growth factor family and are considered related based on their protein function (Hesketh, 1995). The *MYC* genes join this partition along with the *MAX* gene. The human *MYC* genes and the *MAX* gene are also considered to be closely related in terms of protein function (Hesketh, 1995). The MYC, MYC-N and MYC-L proteins are helix-loop-helix/leucine zipper proteins that form sequence-specific DNA binding heterodimers with the *MAX* protein (Hesketh, 1995). Although all of these genes are not members of the same family, they all have some level of involvement as related to fibroblast (spindle-shaped cells responsible for the formation of extracellular fibers such as collagen in connective tissues) growth or transformation (King and Stansfield, 1990, Hesketh, 1995).

Even genes in this partition that are farther apart seem to have closer functional relationships to the genes within the partition than to other genes in the data set. For example, *KRAS2* does not transform normal fibroblasts but transforms transfected (cells that have acquired foreign DNA) fibroblasts when supplemented with oncogenes like *MYC*. *PIM-1* also has this relationship with *MYC*. *WNT-1* is known to transform fibroblasts, but with low efficiency. However, it is known to play a role in the fibroblast transformation process through a paracrine mechanism. Overexpression of *FOS* does not transform human fibroblasts, but rat *FOS* does transform rat fibroblasts under certain conditions . There is no currently known relationship between *WNT-1* and *FOS* (Hesketh, 1995).

51

Ward's and beta-flexible both recommended four clusters rather than three clusters. The preponderance of evidence from the other methods did not support four clusters, but the possibility of four clusters should not be completely eliminated from consideration. Four clusters would formally break cluster 1 into the two partitions discussed above. It is, however, equally possible that Ward's is showing a bias towards clusters of equal size. Another item of interest is that single linkage cluster analysis seems to digress from the other methods. This result supports Milligan's (1995) observations concerning single linkage clustering based on his simulation studies; that single linkage consistently performs poorly.

Regardless of slight variation in the optimal number of clusters, the biological information of interest for this variable remains unchanged. There appears to be a relationship between number of introns in a gene and the gene's regulatory status.

Clustering Based on Number of Introns and Mean Intronic Log Lengths in a Gene

The clustering process based on both the number of introns in a gene and the mean intronic log lengths in a gene is evaluated for the nine clustering methods offered by SAS. Ties did not occur in the SAS output for this combination of variables. The pseudo F and pseudo t^2 results are reported in table 4.5. There is a strong consensus that 10 clusters is optimal.

Table 4.5: Optimal Number of Clusters Based on Number of Introns and Means of Intronic Log Lengths for 33 Genes

Method	Pseudo F	Pseudo t^2
Ward's	10	2,4,8,**10**,12
Beta-Flexible	5,**10**	4,8,**10**,14
Average	5,**10**	3,4,7,13
Centroid	5,9,11	3,4,7,11,12
Complete	10	4,7,**10**,12
EML	10	3,4,7,13
McQuitty's	5,**10**	4,7,**10**,13
Median	3,5,8	3,6,11,13
Single	8	2,8,11

Clusterings based on Ward's, beta-flexible and average linkage are shown in figures 4.4, 4.5, and 4.6 respectively. Each method places the same genes in the ten clusters. Thus, there is 100 percent consensus as to cluster members based on the three preferred methods.
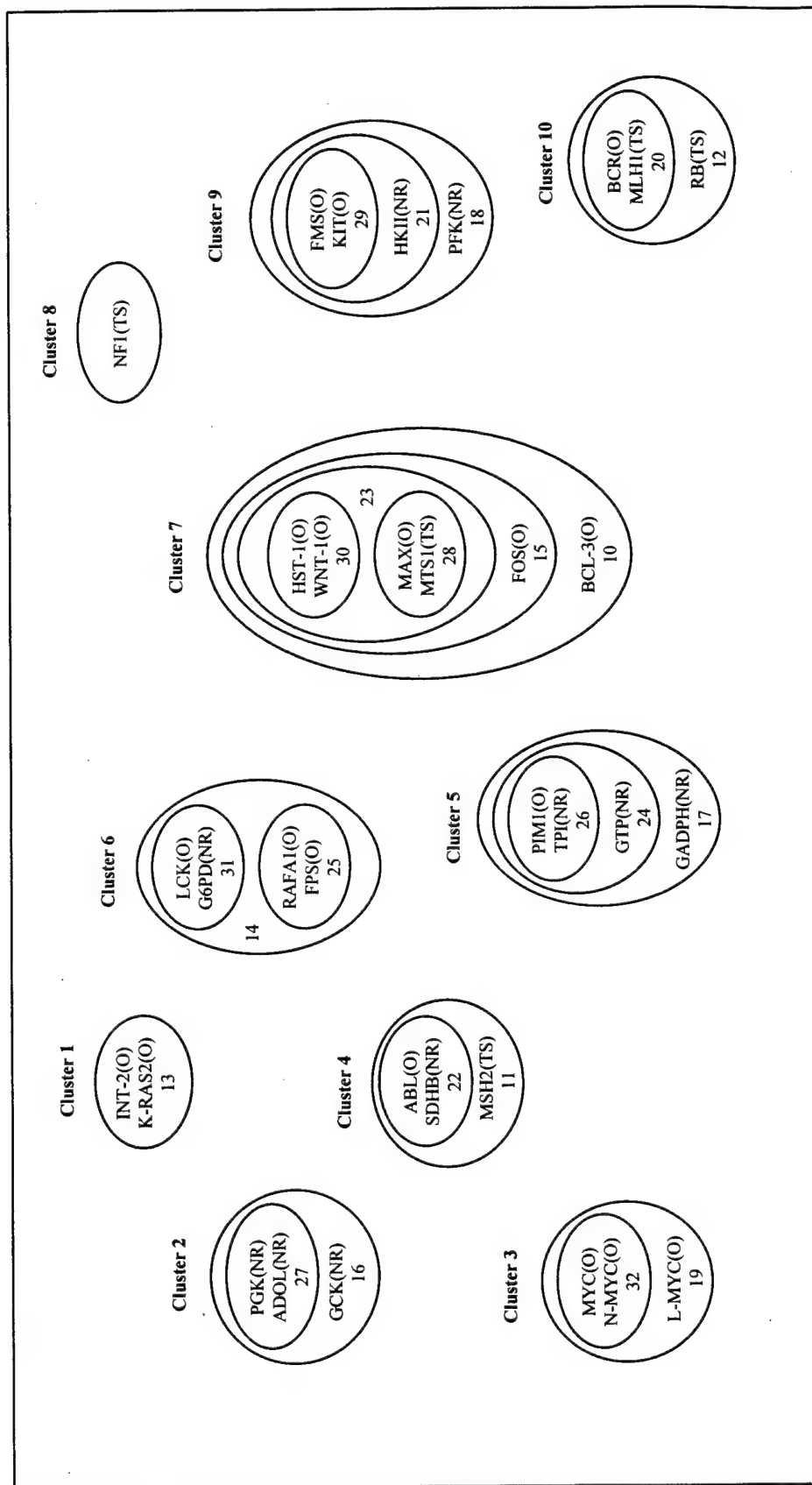
Figure 4.4: Clusters Based on Number of Introns and Means of Intronic Log Lengths Using Ward's Minimum Variance Cluster Analysis
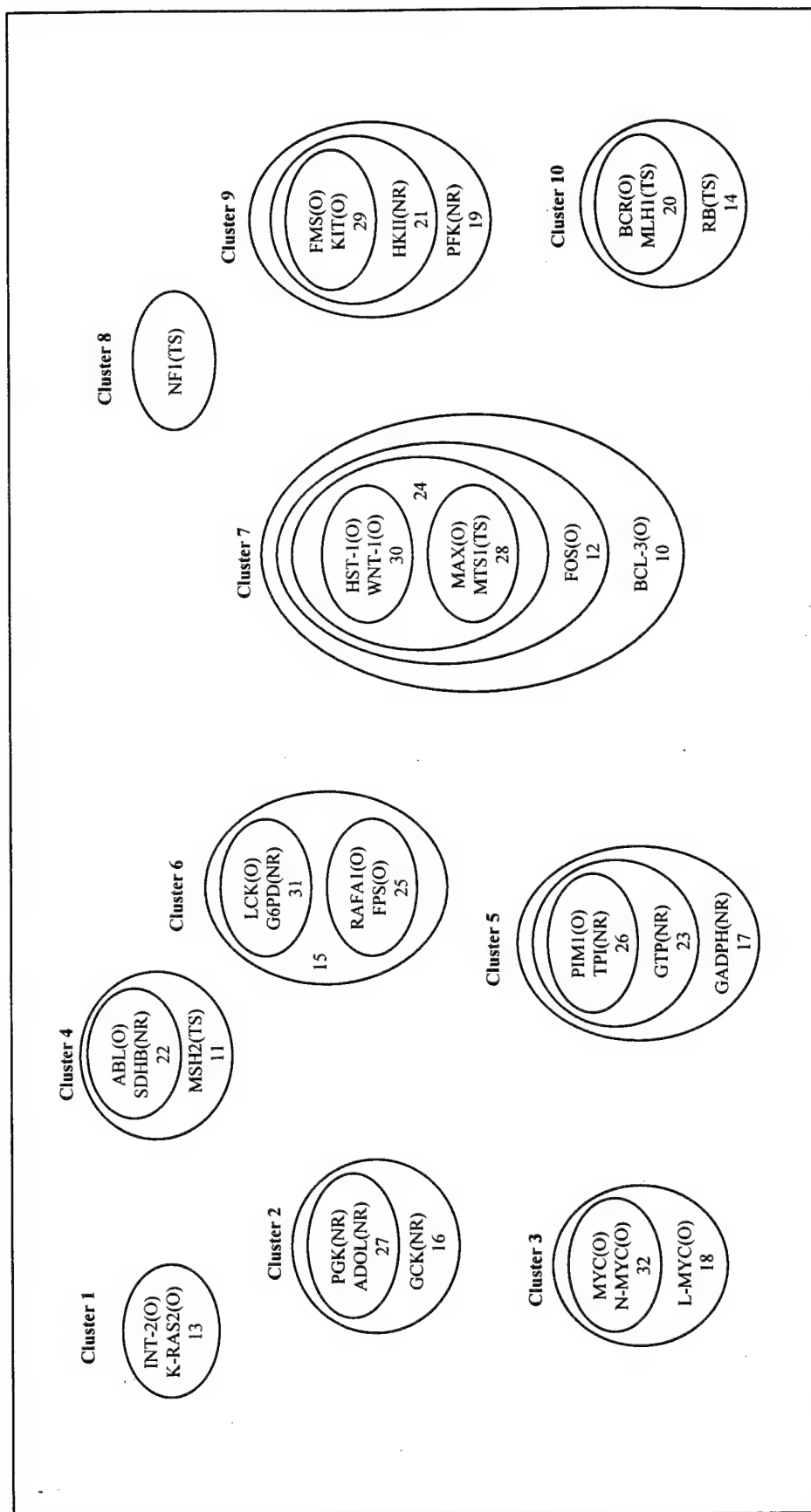
Figure 4.5: Clusters Based on Number of Introns and Means of Intronic Log Lengths Using Beta-Flexible Cluster Analysis
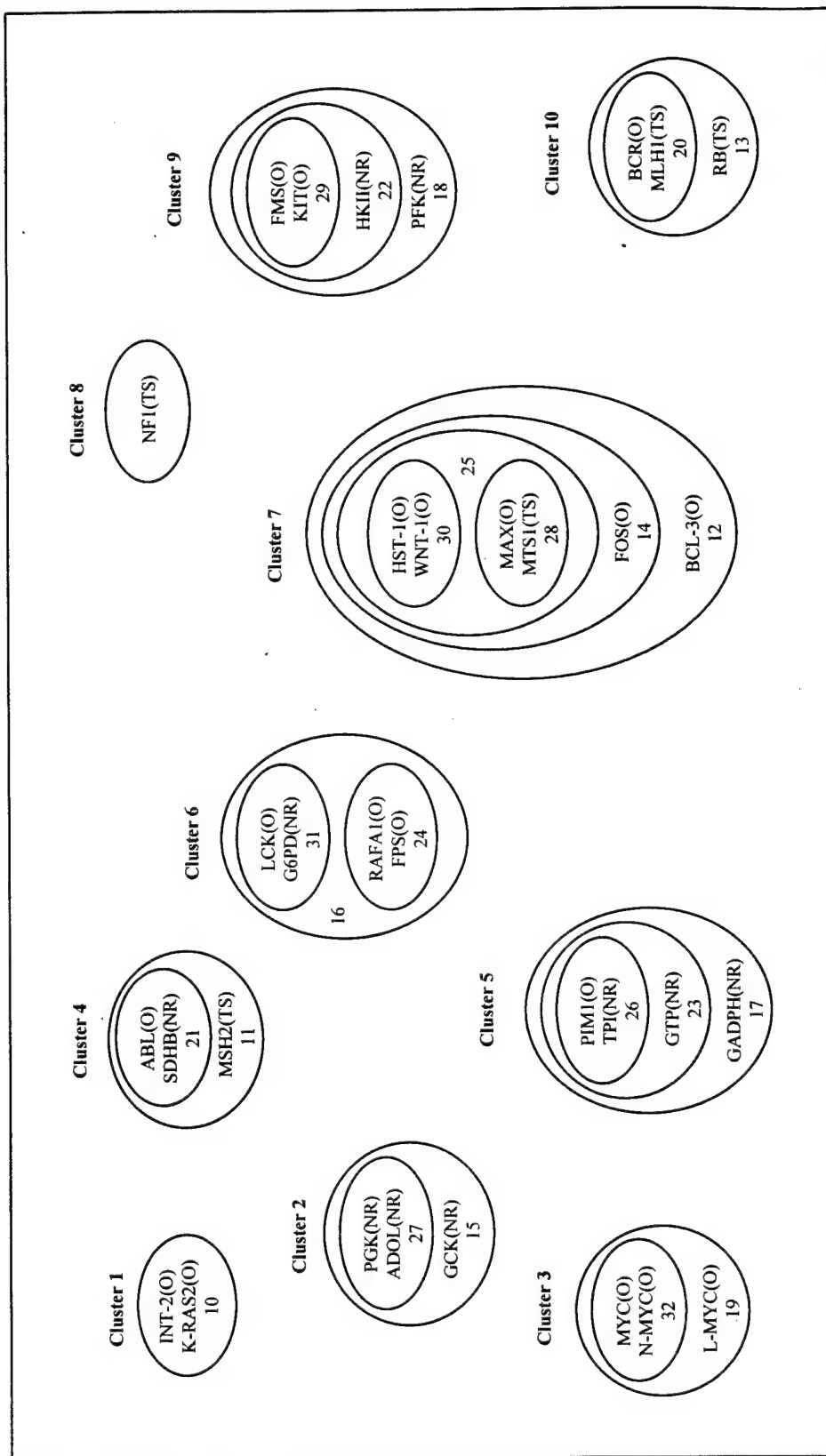
Figure 4.6: Clusters Based on Number of Introns and Means of Intronic Log Lengths Using Average Linkage Cluster Analysis

Interpretation of Clustering

The three *MYC* genes cluster together. *MYC* and *N-MYC* cluster very early in the process, with *L-MYC* joining later. The early clustering of *HST-1* and *WNT-1* is notable since *HST-1* is known to be a fibroblast growth factor and *WNT-1* is also believed to be a growth factor with an unknown receptor. Proteins from both of these genes may have a paracrine mechanism. However, *INT-2* is also a member of the fibroblast growth factor family and it does not cluster with these two. It instead clusters with *K-ras* later in the clustering hierarchy. This may be biologically significant because both *INT-2* and *K-ras* play a role in the transformation of NIH 3T3 cells. NIH 3T3 cells transformed by mouse *INT-2* cDNA express a series of *INT-2*-related proteins; no mention was made as to whether this is also true in humans (Hesketh, 1995). Normal fibroblasts are not transformed by *RAS* oncogenes, but NIH 3T3 fibroblasts are transformed by overexpression of normal *RAS* proteins. Thus, it is biologically plausible that these two genes may share a more distant relationship.

The early clustering of *KIT* and *FMS* is also of interest, especially since these have previously been identified as being structurally similar (Hesketh, 1995). *RAFA1* is a member of the *SRC* super-family of protein kinases. *FPS* is known to have C-terminal homology with *SRC* (Hesketh, 1995). *LCK* is also a member of the *SRC* family and joins with *FPS* and *RAFA1* later in the clustering process. These two are not as closely clustered as some of the other gene relationships, but it may be that they share a more distant relationship. There is also a small cluster of only non-regulatory genes, and several small clusters containing almost exclusively oncogenes.

Conversely, oncogene *LCK* and non-regulatory gene *G6PD* cluster closely for no apparent biological reason. The same is true of *MAX* and *MTS1*. Although *MTS1* is a multiple tumor suppressor and could therefore influence many genes, there is not apparent reason why it would cluster more closely with *MAX* (which involves NIH 3T3 fibroblasts, HeLa cells, neuroblastoma-derived cell lines) than other oncogenes. This is additionally complicated by the fact that *MAX* is known to be related to the *MYC* family and yet it did not cluster with this family. *MAX* did, however, cluster with *WNT-1* , *HST-1, FOS and BCL-3*. With the exception of *BCL-3*, all of these have relationships that have previously been discussed. Little is known about *BCL-3*, so it is difficult to determine where this gene belongs.

In conclusion, these clusters suggest that there may be a relationship based on number of introns and intronic log lengths of the introns. There are some genes that are not close when observing clusters based on the two variables separately, but seem to move closer when these two variables combine. It is also interesting that the optimal number of clusters indicates more, smaller groups.

It may be that we would see more (or fewer) relationships if more data were available. Meanwhile, the clustering provides at least partial evidence of non-randomness.

CHAPTER V

TEST FOR INDEPENDENCE VERSUS A FIRST-ORDER MARKOV PROCESS

Methodology

DNA can be considered as a stochastic process $X_1,...,X_n$, where each observation

assumes one of four states: *A, G, C,* or *T* corresponding to the four nucleotides. If $X_m = i$,

the process is said to be in state *i* at time *m*. Whenever the process is in state *i*, there is a

fixed probability p_{ij} that the process will next be in state *j*. This is expressed as

$$P\{X_{m+1} = j | X_m = i\} = p_{ij} \qquad \text{(Ross, 1989)}$$

for all states *i, j* and all $m \geq 0$. In this situation, the DNA sequence may be described by

a first-order Markov chain model. In words, the above equation states that, for a Markov

chain, the conditional distribution of any state X_{m+1} given the past states

$X_0,\ X_1,\ \ldots,X_{m-1}$, and the present state X_m, is independent of the past states and depends

only upon the present state. p_{ij} is one-step transition matrix and the Markov chain it

represents is said to be of first-order. Properties of p_{ij} are

$$p_{ij} \geq 0 \ \text{where } i \text{ and } j \text{ are } \geq 0, \ \text{and} \ \sum_{j=1}^{s} p_{ij} = 1, \text{where } i = 1, 2, \ldots, \quad \text{(Ross, 1989)}$$

where *s* is the number of states. A Markov chain is said to have order zero if it is a

sequence of independent random variables, i.e., p_{ij} does not depend upon *i*.

The idea that a DNA sequence (especially the intronic regions) might be represented

by a Markov chain has been discussed in the literature without resolution. It seems

appropriate to use a statistical hypothesis testing methodology to investigate this issue.

Billingsley (1961) provides a Chi-square test for a null hypothesis that the data are independent (i.e., a zero-order Markov process) against the alternative hypothesis that the data come from a first-order Markov process. This test will be applied to the data in this thesis to investigate whether introns within a gene show evidence of independence or a first-order Markov process. An examination of introns within a gene has not previously been investigated in the literature and it seems appropriate to test for evidence of a first-order Markov process before performing the test of homology in Chapter VI.

To test whether an observed sequence is really an independent sequence, let $\{x_1, x_2, ...,$ $x_n\}$ be a sample from a first-order Markov process with transition probabilities p_{ij}. The $s \times s$ matrix $F = \{f_{ij}\}$ provides the transition count of the sequence, i.e., the number of transitions from state i to state j. The hypotheses are

$H_0 : p_{ij} = p_j$ is independent of i for all i versus $H_1 : p_{ij}$ is not independent of i for some i.

The probability of obtaining a particular sequence begins at x_1 and has a transition count matrix F. Let

$$ f_{i.} = \sum_j f_{ij} \quad \text{and} \quad f_{.j} = \sum_i f_{ij} . $$

$\{f_{i.}\}$ is the number of transitions out of state i (where $f_{i.} = f_{.i} = f_i$) and $\{f_{.j}\}$ is the number of transitions into state j. This means

$$ \sum_{ij} f_{ij} = \sum_i f_i = \sum_j f_{.j} = n - 1 . $$

The following statistic can now be utilized

$$\sum_{ij} \frac{\left(f_{ij} - f_{i\cdot}f_{\cdot j} / (n-1)\right)^2}{f_{i\cdot}f_{\cdot j} / (n-1)},$$ (Billingsley, 1961)

which is chi-square with $s(s-1)-(s-1) = (s-1)^2$ degrees of freedom for a first-order Markov

process.

It can be shown that this chi-square statistic is approximately a Neyman-Pearson

likelihood ratio test

$$\sum_{ij} \frac{\left(f_{ij} - f_{i\cdot}f_{\cdot j} / (n-1)\right)^2}{f_{i\cdot}f_{\cdot j} / (n-1)} \sim 2\sum_{ij} f_{ij} \ln \frac{f_{ij}}{f_{i\cdot}f_{\cdot j} / (n-1)}.$$ (Billingsley, 1961)

In the limit, each formula has a Chi-square distribution with $(s-1)^2$ degrees of freedom.

Thus, the Chi-square test for zero-order versus first-order in a Markov process is

asymptotically equivalent to performing a likelihood ratio test based on multinomial

distributions.

Application

As an example, the first-order transition counts for the first intron of *BCL*-3 is shown

in table 5.1. The length of the first intron of *BCL-3* is 1355 bp. Matrix *F* is provided as

follows:

Table 5.1: First-order Transition Count for Intron 1 of BCL-3

	A	G	C	T	
A	76	143	61	73	353
G	127	194	54	70	445
C	105	13	100	73	291
T	45	95	76	49	265
	353	445	291	265	1354

The Chi-square test statistic is

$$\chi^2 = \sum_{ij} \frac{\left(f_{ij} - f_i f_{\cdot j} / (n-1)\right)^2}{f_i f_{\cdot j} / (n-1)} = 175.51,$$

which is much larger than $\chi^2_9 = 21.666$ (where $\alpha = 0.01$ and the degrees of freedom are computed from $(s-1)^2 = (4-1)^2 = 9$). A small α was chosen to provide a stringent hypothesis test. Thus, the null hypothesis of independence is rejected and it is concluded that the intron 1 of *BCL-3* sequence more consistent with a first-order Markov process than with independence.

An *F* transition matrix is created for all 375 introns in the 33 genes. Not all intronic information is available for every intron in every gene. Unfortunately, with intronic regions, the actual location of the missing data is often unknown. Table B.1 depicts the amount of intronic sequence data available. The S-Plus code is provided in appendix E.

Table 5.2 depicts the results for the 375 introns in the data set. White cells represent independent intronic regions. Light gray cells represent intronic sequences showing evidence of a first-order Markov process.

Sixty-seven percent of the introns show evidence of a first-order Markov process. It is very interesting to note that a given gene may have varying degrees of independence. This may explain why, in the biological literature, there are mixed results as to whether there is evidence of a Markov process in sequences. From only a random sample, it is possible to obtain mixed results.

Table 5.2: Results From Test of Independence Versus Markov Structure For Introns in Each Gene
(Gray Cells Reject at the α = .01 Level)

Gene	# of Introns In Gene	Introns															
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
ABL(O)	10	5845.08	31.70	398.28	618.49	35.61	129.23	160.33	55.09	34.74	215.59						
BCR(O)	22	5490.28	724.30	29.14	516.86	310.45	164.03	67.02	1105.58	93.76	215.04	108.88	179.16	88.53	188.13	186.69	1188.73
FMS(O)	21	101.49	543.90	47.76	195.96	90.21	315.05	213.49	25.39	29.07	192.29	665.79	26.41	97.21	107.68	197.88	25.75
BCL-3(O)	8	175.51	117.58	227.75	7.86	15.65	12.82	34.95	215.21								
FOS(O)	3	32.92	43.74	16.10													
HST-1(O)	2	36.60	84.35														
INT-2(O)	2	93.33	740.97														
LCK(O)	11	52.34	19.04	20.67	8.81	16.82	35.41	29.42	55.01	19.34	14.08	34.51					
MYC(O)	2	85.91	129.05														
L-MYC(O)	2	20.86	271.08														
N-MYC(O)	2	77.33	182.29														
MAX(O)	1	77.92															
PIM1(O)	5	8.91	8.89	23.07	130.63	75.68											
KIT(O)	20	8.78	18.94	20.80	24.17	18.09	12.61	18.45	34.97	35.43	9.06	34.92	18.77	21.05	23.50	30.79	34.78
RAFA1(O)	15	93.29	22.36	89.81	21.61	22.78	72.59	16.32	25.14	13.62	94.69	25.13	58.41	46.08	99.11	80.02	
FPS(O)	18	64.55	38.55	118.56	20.76	168.87	14.16	32.36	25.16	10.75	73.44	75.44	55.07	57.31	23.73	27.37	37.05
WNT-1(O)	3	44.97	44.17	33.74													
K-RAS2(O)	4	23.50	38.47	28.74	57.05												
MLH1(TS)	18	12.10	20.00	22.83	15.46	23.73	22.51	12.99	35.15	43.89	28.96	25.13	29.93	36.71	42.47	9.24	18.25
MSH2(TS)	15	12.99	20.90	15.29	11.04	22.36	24.70	14.23	16.58	38.78	16.40	35.61	14.29	24.74	9.82	22.86	
MTS1(TS)	3	39.32	108.51	70.21													
RB(TS)	26	16.93	26.68	21.64	13.87	21.07	23.45	21.20	27.97	19.60	21.06	28.14	7.99	37.03	51.20	24.60	32.82
NF1(TS)	56	12.74	13.37	15.70	33.60	13.44	14.42	12.19	10.23	17.59	8.20	71.73	5.12	19.40	25.46	12.45	20.29
G6PD(NR)	12	26.26	786.38	30.03	79.38	74.45	36.73	52.46	91.31	26.80	24.87	24.38	22.01				
PGK(NR)	10	23.50	13.71	34.47	25.48	25.32	19.40	25.32	28.20	16.97	23.31						
ADOL(NR)	8	286.91	80.88	87.40	53.32	54.57	29.47	54.90	171.87								
GTP(NR)	8	48.05	32.60	17.12	30.70	15.83	17.96	40.00	47.15								
PFK(NR)	21	18.54	14.09	23.03	43.48	29.92	18.95	158.57	29.58	15.60	55.43	16.33	16.89	21.71	7.81	20.54	8.83
TPI(NR)	6	94.96	22.65	21.47	48.43	43.92	26.50										
GCK(NR)	9	250.83	48.40	53.15	57.24	37.69	29.82	36.91	20.94	32.89							
GAPDH(NR)	8	5.92	121.51	18.25	34.53	13.72	13.55	33.53	29.44								
SDHB(NR)	7	60.17	8.54	87.18	40.01	27.18	42.33	50.49									
HKII(NR)	17	31.61	29.47	17.13	31.90	52.74	20.77	9.90	39.90	24.01	27.13	25.56	13.16	30.47	20.53	20.31	22.47

Table 5.2 (Continued): Results for Test of Independence Versus Markov Structure For Introns in Each Gene (Gray Cells Reject at the $\alpha = .01$ Level)

Gene	Introns																			
	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36
ABL(O)	82.09	117.23	132.06	158.53	95.53	100.69														
BCR(O)	70.86	15.78	75.65	77.10	15.29															
FMS(O)																				
BCL-3(O)																				
FOS(O)																				
HST-1(O)																				
INT-2(O)																				
LCK(O)																				
MYC(O)																				
L-MYC(O)																				
N-MYC(O)																				
MAX(O)																				
PIM1(O)																				
KIT(O)	12.97	21.40	28.81	27.01																
RAFA1(O)	27.81	133.84																		
FPS(O)																				
WNT-1(O)																				
K-RAS2(O)																				
MLH1(TS)	7.71	34.28																		
MSH2(TS)																				
MTS1(TS)																				
RB(TS)	13.31	24.75	30.10	67.50	19.20	13.64	17.52	20.18	27.65	19.46										
NF1(TS)	12.58	27.59	16.87	16.16	25.59	12.30	26.26	42.98	15.26	11.96	9.25	14.24	26.05	72.77	19.08	57.72	13.35	24.56	1554.67	49.95
G6PD(NR)																				
PGK(NR)																				
ADOL(NR)																				
GTP(NR)																				
PFK(NR)	16.26	16.58	14.22	12.18	53.54															
TPI(NR)																				
GCK(NR)																				
GAPDH(NR)																				
SDHB(NR)																				
HKII(NR)	17.69																			

Table 5.2(Continued): Results For Test of Independence Versus Markov Structure For Introns in Each Gene (Gray Cells Reject at the α = .01 Level)

Gene	Introns																			
	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56
ABL(O)																				
BCR(O)																				
FMS(O)																				
BCL-3(O)																				
FOS(O)																				
HST-1(O)																				
INT-2(O)																				
LCK(O)																				
MYC(O)																				
L-MYC(O)																				
N-MYC(O)																				
MAX(O)																				
PIM1(O)																				
KIT(O)																				
RAFA1(O)																				
FPS(O)																				
WNT-1(O)																				
K-RAS2(O)																				
MLH1(TS)																				
MSH2(TS)																				
MTS1(TS)																				
RB(TS)																				
NFI(TS)	155.26	180.29	66.89	23.17	16.60	10.58	9.78	27.13	66.44	98.75	303.11	46.00	80.16	202.44	14.18	21.93	55.29	15.12	81.89	664.21
G6PD(NR)																				
PGK(NR)																				
ADOL(NR)																				
GTP(NR)																				
PFK(NR)																				
TPI(NR)																				
GCK(NR)																				
GAPDH(NR)																				
SDHB(NR)																				
HKII(NR)																				

Table 5.3 displays the percentage of intronic regions that show evidence of a first-order

Markov process by gene.

Table 5.3 Percentage of Intronic Regions Showing Evidence of a Markov Process

Gene	Percent (%)
ABL(O)	100
BCR(O)	100
FMS(O)	90.5
BCL-3(O)	62.5
FOS(O)	66.7
HST-1(O)	100
INT-2(O)	100
LCK(O)	45.5
MYC(O)	100
L-MYC(O)	50
N-MYC(O)	100
MAX(O)	100
PIM1(O)	60
KIT(O)	45
RAFA1(O)	80.0
FPS(O)	83.3
WNT-1(O)	100.0
K-RAS2(O)	100.0
MLH1(TS)	66.7
MSH2(TS)	40.0
MTS1(TS)	100.0
RB(TS)	46.2
NF1(TS)	50.0
G6PD(NR)	100.0
PGK(NR)	70.0
ADOL(NR)	100.0
GTP(NR)	62.5
PFK(NR)	38.1
TPI(NR)	83.3
GCK(NR)	88.9
GAPDH(NR)	50.0
SDHB(NR)	85.7
HKII(NR)	58.8

Two additional issues of interest are: (1) the power to detect statistically significant differences, and (2) multiple testing. To investigate whether all the independent sequences were only the shortest ones, I superimposed the gray cells over the intronic lengths. This is depicted in table B.1. There are some short sequences that are independent, but there are also some very long sequences that are clearly independent. Similarly there are some fairly short sequences that show evidence of a Markov process. Although there is a small sample power issue here, it does not seem to be a severe problem.

For the multiple testing problem, when testing at the 1 percent level, one expects by chance alone to have 1 percent of the cells be significant. Clearly, there are many more cells that reject the null hypothesis. There really was no way to avoid this problem. Since the intronic regions are processed individually in the body, it would not be appropriate to link them together (as is often done with the exonic regions). This means that it is not possible to do a hypothesis test of the whole DNA sequence before testing the individual intronic subsequences. This is the approach recommended in Zerbe and Murphy (1986), but it cannot be incorporated herein. Thus, it must be acknowledged that about intronic regions may show a first-order Markov process simply by chance.

Now, having sufficient evidence of Markov process for a majority of introns, I can proceed to test the homology (similarity) of the introns.

CHAPTER VI

CLUSTER ANALYSIS INCORPORATING
A MARKOV HOMOLOGY TEST

Methodology

Test of Homology

In the last chapter, a test of sequential independence for one sample (i.e., one intron) was introduced. In this chapter, the case of comparing two samples is investigated. Billingsley refers to this methodology as a test of homology (Billingsley, 1961) for two first-order Markov processes. In this case, $\{f_{ij}\}$ is the transition count of one sample and $\{g_{ij}\}$ is the transition count of a second sample. The two samples are independent of each other and have Markov chains with transition probabilities p_{ij} and q_{ij}, estimated by f_{ij}/f_i and g_{ij}/g_i respectively. If $p_{ij} = q_{ij}$, for all i,j, then the common estimate is $(f_{ij} + g_{ij}) / (f_i + g_i)$. Using this estimate, it is possible to create a chi-square test statistic for a test of homology (similarity). Hypotheses are

$$H_0: p_{ij} = q_{ij} \text{ for all } i,j \quad \text{versus} \quad H_1: p_{ij} \neq q_{ij} \text{ for some } i,j$$

The chi-square test statistic is

$$\sum_{ij} \frac{\left[f_{ij} - f_i \left(\frac{f_{ij} + g_{ij}}{f_i + g_i} \right) \right]^2}{f_i \frac{f_{ij} + g_{ij}}{f_i + g_i}} + \sum_{ij} \frac{\left[g_{ij} - g_i \left(\frac{f_{ij} + g_{ij}}{f_i + g_i} \right) \right]^2}{g_i \frac{f_{ij} + g_{ij}}{f_i + g_i}}$$

$$= \sum_{ij} \frac{f_i g_i}{f_{ij} + g_{ij}} \left(\frac{f_{ij}}{f_i} - \frac{g_{ij}}{g_i} \right)^2 . \qquad \text{(Billingsley, 1961)}$$

The asymptotic distribution has $s(s-1)$ degrees of freedom as $n \rightarrow \infty$, where s (the number of states) is fixed. This formula can also be extended to n samples. In this case, the degrees of freedom is $(n-1)s(s-1)$.

Combining Chi-square Test Statistics

Having computed the Chi-square test statistics for all corresponding intron pairs between two genes, the mean of the Chi-square test statistics is computed. This provides a global relationship for the intronic regions between two genes. This mean becomes a measure of distance that is used in a distance matrix to perform a cluster analysis. This distance matrix is taken directly from the relationships in the data and is an example of a one-mode structure. There is no evidence in the literature of a test statistic being used as a distance measure. This type of distance measure may provide informative clustering information than using features information.

Billingsley's homology statistic is applied to the corresponding intronic regions of DNA from each gene pair. Corresponding intronic regions were chosen because, *in vivo*, the introns are transcribed in order in the nucleus. It is appropriate to compare intronic region 1 from gene 1 with intronic region 1 from gene 2, and then intronic region 2 from gene 1 with intronic region 2 from gene 2, etc., until all paired intronic regions have been completed out to the number of introns in the gene which has the fewest.

This approach differs from the usual scores based on percent match found in DNA analysis. All the available intronic information is used to compute the transition matrices for each region. There is no truncation of the DNA data from an intronic region. We are observing ordered trend for all the data in the DNA sequences rather than alignment. This seems appropriate in light of the fact that intronic regions are known to be less

69

conserved and have more gaps. The ordered relationships will generally not be as apparent as those found in the coding regions. This is a different way to determine statistical homology.

The advantage here is that we have an actual hypothesis testing methodology which is lacking in the current sequence analysis methodologies. The hypotheses are

H_0: $p_{ij} = q_{ij}$, for all i,j (i.e., the two intronic regions are homologous)

versus

H_1: $p_{ij} \neq q_{ij}$, for some i,j

where p_{ij} and q_{ij} are the transition probability matrices for each pair of intronic regions from each gene pair.

Application

As an example, suppose intron 1 from *BCL*-3 is again chosen. The transition count is provided below in table 6.1. Intron 1 from *BCL*-3 is compared with intron 1 from *FOS*. The transition counts for *FOS* are given in table 6.2.

Table 6.1: First-Order Transition Count for *BCL*-3, Intron 1

	A(1)	G(2)	C(3)	T(4)
A(1)	76	143	61	73
G(2)	127	194	54	70
C(3)	105	13	100	73
T(4)	45	95	76	49

Table 6.2: First- Order Transition Count for *FOS*, Intron 1

	A(1)	G(2)	C(3)	T(4)
A(1)	34	60	31	24
G(2)	58	100	69	33
C(3)	41	50	51	54
T(4)	16	50	45	36

Since the corresponding transition count matrices have dimension 4 x 4, the degrees of freedom will be $s(s-1) = 4(4-1) = 12$. When $\alpha = 0.01$, the critical point is 26.217.

$$\chi^2 = \sum_{ij} \frac{f_i g_i}{f_{ij} + g_{ij}} \left(\frac{f_{ij}}{f_i} - \frac{g_{ij}}{g_i} \right)^2 = 82.15092 \text{ for } BCL\text{-}3 \text{ and } FOS \text{ introns } 1.$$

The null hypothesis is rejected; these two introns are not homologous.

This hypothesis test is performed for all possible corresponding intron pairs. The S-Plus function to compute the Chi-square test statistic is provided in appendix E. Results, by gene, are provided in tables C.1 - C.33; white cells denote statistical homology while gray cells indicate a lack of homology.

Some interesting results can be noted in these tables. Two genes may have some intronic regions which are statistically different while others show evidence of homology. Numerous gene pairs are completely non-homologous, while no two genes show evidence of complete homology.

Of even greater interest is the fact that many of the oncogenes appear to have more similarity with several non-regulatory genes than other oncogenes. This will be further examined after the clustering process.

The clustering process described in Chapter IV is performed, with several subtle differences. Since I provide my own distance matrix which is already squared, the

nosquare option in SAS is chosen. The SAS code for this analysis is provided in appendix D. When one uses a one-mode structure, the ability to use the EML clustering process is lost. This is because EML relies on the number of variables (i.e., coordinate data) for its calculation. This information is not available when a unique distance matrix is incorporated. For the same reason, the ability to compute the pseudo F and pseudo t^2 is lost for all methods except Ward's, average linkage, and centroid clustering. Results from the pseudo F and pseudo t^2 for these methods are shown in table 6.3. This means that there is less information from which to draw a consensus as to the optimal number of clusters. From the information remaining, there appears to be a consensus that the optimal number of clusters is 2.

Table 6.3: Optimal Number of Clusters Based on Mean Chi-square Homology Statistics

Method	Pseudo F	Pseudo t^2
Ward's	2	2,3,8
Flexible-Beta	NA	NA
Average	2,7,14	2,7,10,12
Centroid	2,4,6,9,14	2,5,12,14
Complete	NA	NA
EML		
McQuitty's	NA	NA
Median	NA	NA
Single	NA	NA

Clusters based on Ward's, flexible-beta and average linkage are shown in figures 6.1, 6.2, and 6.3 respectively. As there is not 100 percent consensus between methods as to cluster members, table 6.4 summarizes the cluster members by method.
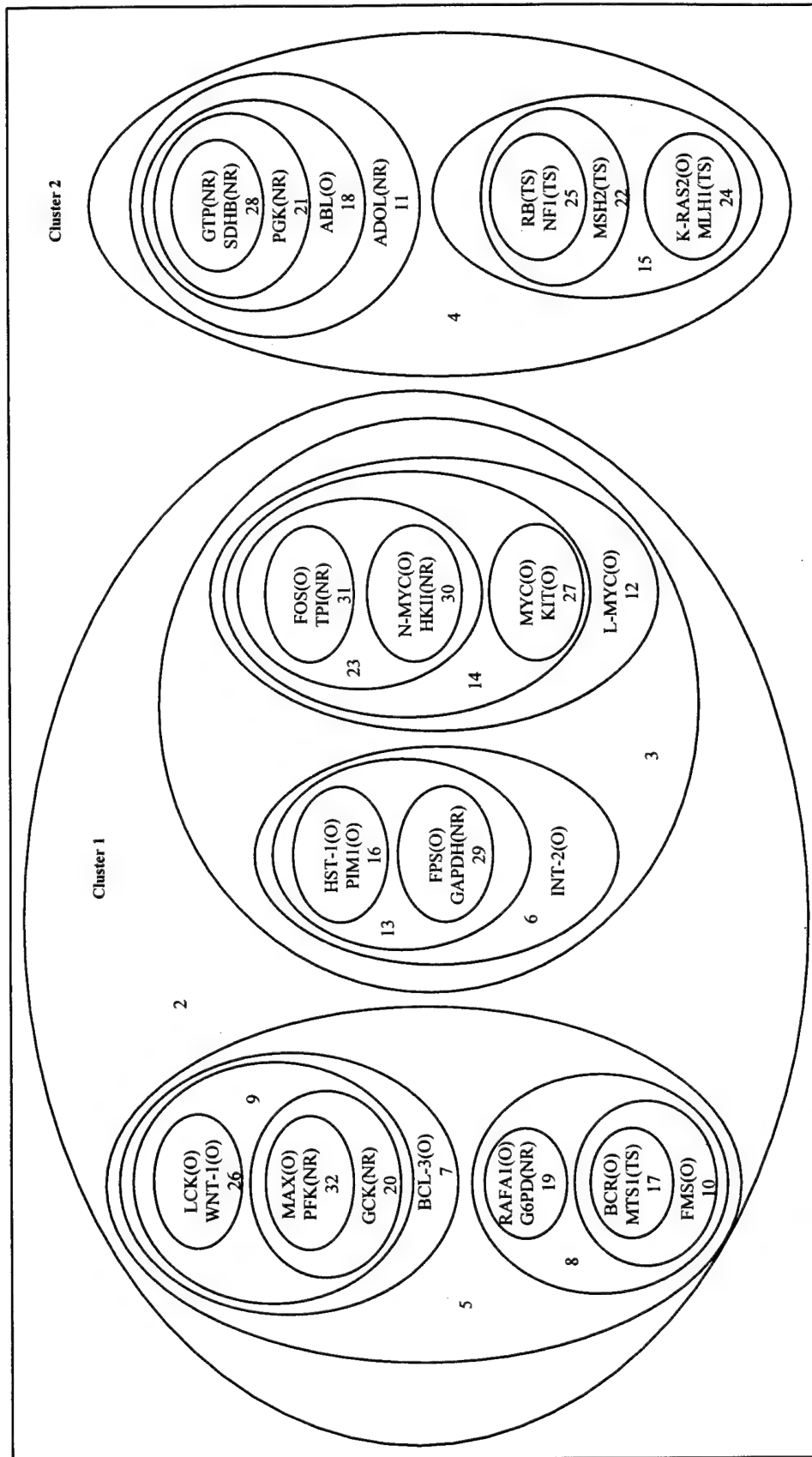
Figure 6.1: Clusters Based on Homology Chi-Square Test Results Using Ward's Minimum Variance Cluster Analysis
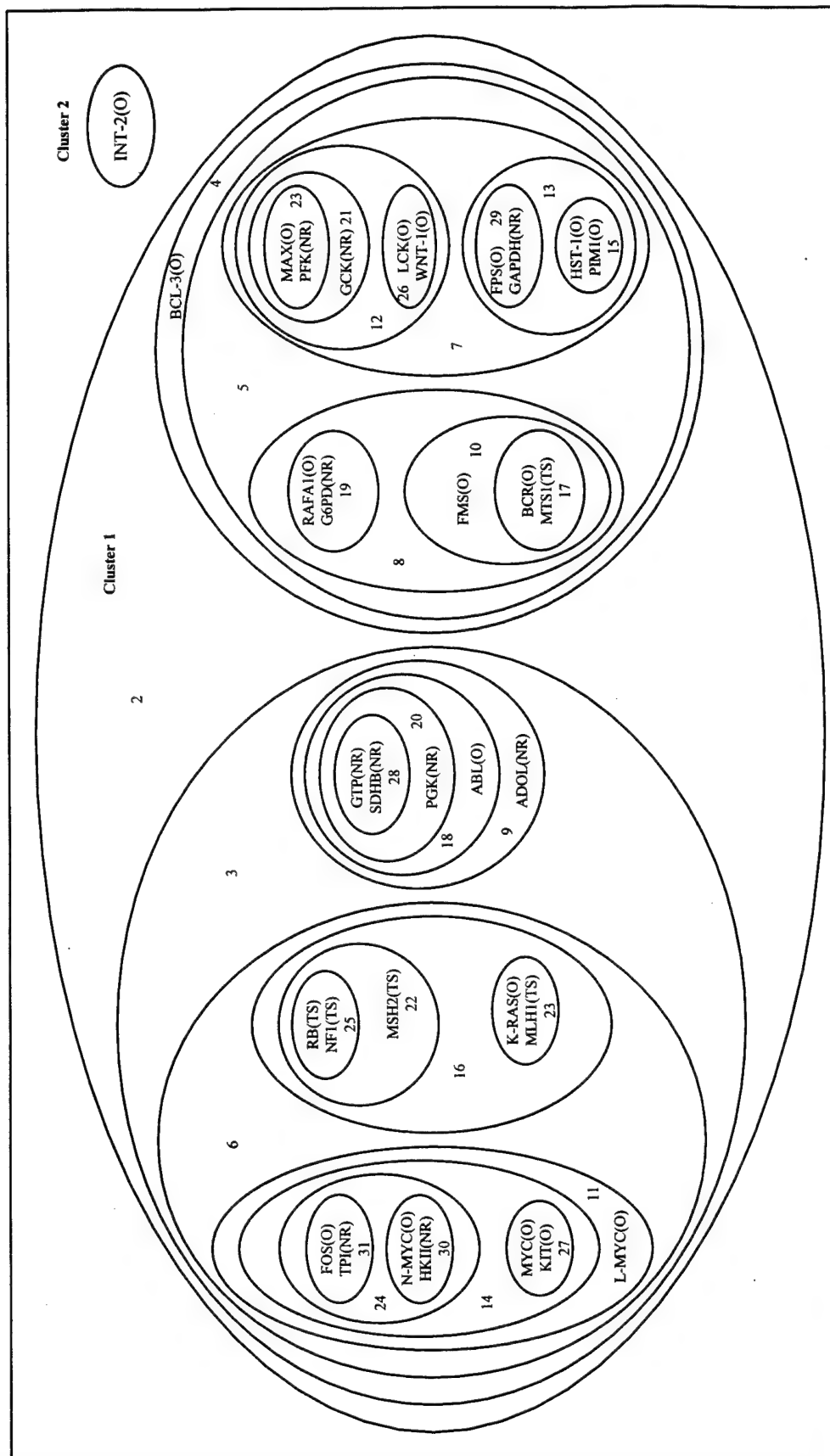
Figure 6.2: Clusters Based on Homology Chi-Square Test Results Using Beta-Flexible Cluster Analysis
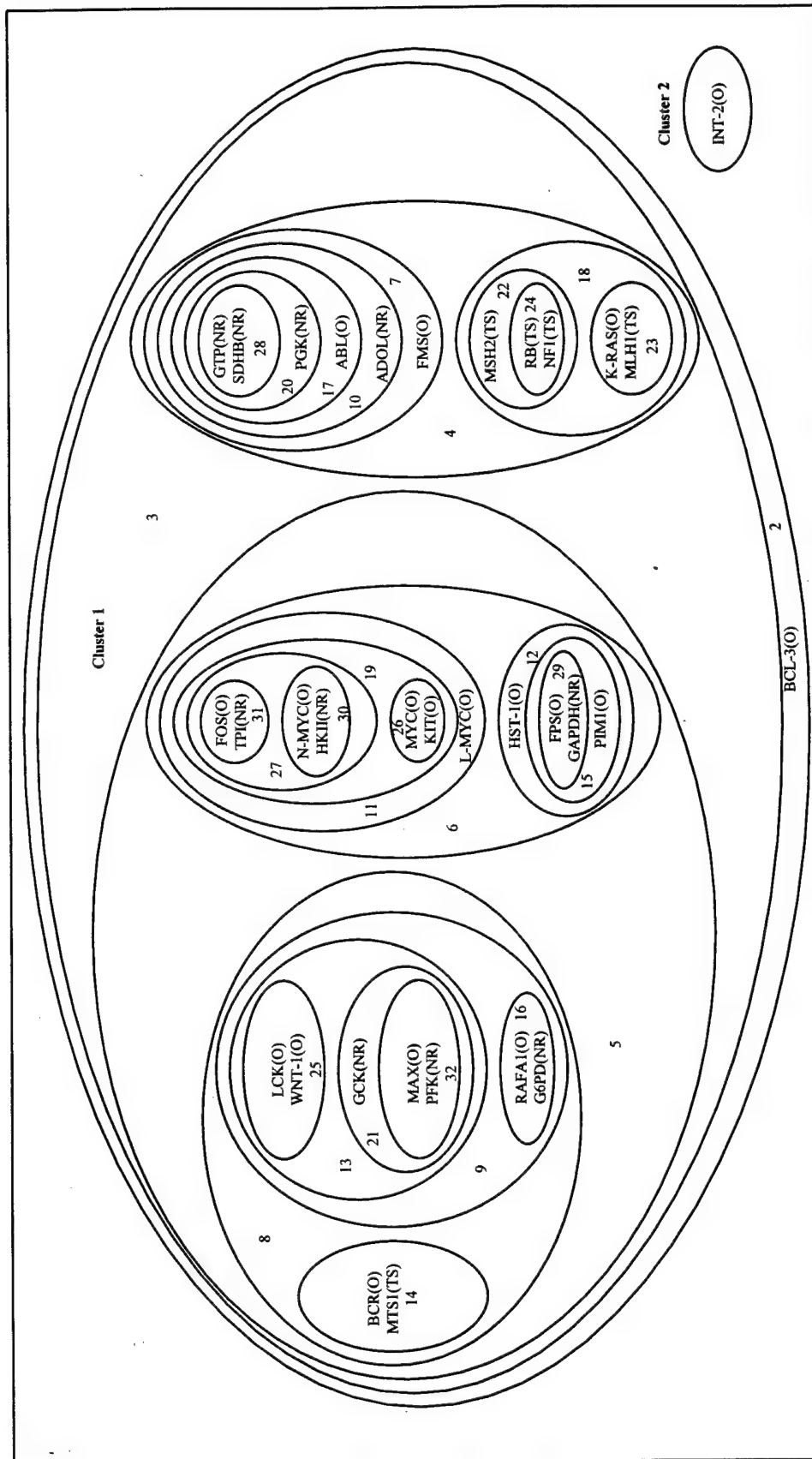
Figure 6.3: Clusters Based on Homology Chi-Square Test Results Using Average Linkage Cluster Analysis

Table 6.4: Comparison of Cluster Members for Preferred Methods Based on Mean Chi-Square Homology Results After Comparing Introns Between Genes

Cluster	Ward's	Average	Beta-Flexible
	LCK(O)	LCK(O)	LCK(O)
	WNT-1(O)	WNT-1(O)	WNT-1(O)
	MAX(O)	MAX(O)	MAX(O)
	RAFA1(O)	RAFA1(O)	RAFA1(O)
	BCR(O)	BCR(O)	BCR(O)
	FMS(O)	FMS(O)	FMS(O)
	HST-1(O)	HST-1(O)	HST-1(O)
	PIM1(O)	PIM1(O)	PIM1(O)
	FPS(O)	FPS(O)	FPS(O)
	BCL-3(O)	BCL-3(O)	BCL-3(O)
	FOS(O)	FOS(O)	FOS(O)
Cluster 1	N-MYC(O)	N-MYC(O)	N-MYC(O)
	MYC(O)	MYC(O)	MYC(O)
	L-MYC(O)	L-MYC(O)	L-MYC(O)
	KIT(O)	KIT(O)	KIT(O)
		ABL(O)	ABL(O)
		K-RAS2(O)	K-RAS2(O)
	INT-2(O)		
		NF1(TS)	NF1(TS)
		MSH2(TS)	MSH2(TS)
		MLH1(TS)	MLH1(TS)
	MTS1(TS)	MTS1(TS)	MTS1(TS)
	G6PD(NR)	G6PD(NR)	G6PD(NR)
	GCK(NR)	GCK(NR)	GCK(NR)
	PFK(NR)	PFK(NR)	PFK(NR)
	HKII(NR)	HKII(NR)	HKII(NR)
	TPI(NR)	TPI(NR)	TPI(NR)
	GAPDH(NR)	GAPDH(NR)	GAPDH(NR)
		GTP(NR)	GTP(NR)
		SDHB(NR)	SDHB(NR)
		PGK(NR)	PGK(NR)
		ADOL(NR)	ADOL(NR)
		INT-2(O)	INT-2(O)
	ABL(O)		
	K-RAS2(O)		
	RB(TS)		
	NF1(TS)		
Cluster 2	MSH2(TS)		
	MLH1(TS)		
	GTP(NR)		
	SDHB(NR)		
	PGK(NR)		
	ADOL(NR)		

Interpretation of Clustering

The beta-flexible method and the average linkage cluster analysis provide similar clusterings. Ward's minimum variance cluster analysis provides slightly different clusters from these two methods. A possible explanation for this difference is the bias Ward's method has demonstrated in simulation studies toward clusters of equal size. For the other two methods, cluster 2 only has one gene, while the cluster 2 based on Ward's method has ten genes. Although there is not complete consensus on cluster members, there are some very informative results. Note that four out of five tumor suppressor genes form a partition within one of the clusters. Despite the lack of consensus concerning the other partition with which the tumor suppressors eventually join, they remain together for all clustering methods. It is also of interest that the *MYC* genes fuse into the same partition, although other genes fuse with them.

There are some non-regulatory genes that group closely with oncogenes. Other non-regulatory genes combine to form an almost totally non-regulatory partition within cluster 2 (only oncogene *ABL* joins them). It is possible that the inclusion of *ABL* is due to a totally random occurrence. However, it is also possible that this result might be biologically meaningful. The tables in appendix C provide many examples of oncogenes which are consistently more similar to a certain non-regulatory gene than to the other oncogenes. The possible biological meaning behind the relationship (if any) between these two gene types is unknown.

CHAPTER VII

CONCLUSIONS

In this chapter, the strengths and weaknesses of analyzing data using cluster analysis are discussed. The biological implications of the clustering results are also reviewed. Finally, some areas for further research are suggested.

Statistical Conclusions

When properly used, cluster analysis allows the determination of possible natural groupings within data. However, cluster analysis is often improperly applied, usually because conclusions are based on only one method (often that which best represents a desired conclusion). Simulation studies can provide recommendations concerning the optimal clustering method for various cluster sizes and shapes. But these simulation studies assume the user knows the sizes and shapes of the clusters in advance and this information is rarely obtainable.

A partial solution is to look for consensus among several methods. If they all concur, there is a greater possibility that a natural ordering occurs within the data. However, this involves a considerable time investment to analyze all possible methods for all possible variable configurations. Cluster analysis is not an expedient process. Additionally, some of the better software packages for this analysis, for example, SAS, do not have the graphics capability to provide visualization of the clusters. This encourages the uninformed user to use software offering weaker clustering methods but better graphics, for example, S-Plus.

SAS provided statistical tools (i.e., the pseudo F and pseudo t^2) for determining the optimal number of clusters. This was of considerable benefit because these two tests had performed admirably in Milligan's simulation studies. But neither is a perfected analytic technique. No critical value is provided that conclusively resolves the proper number of clusters issue for the researcher. Unless pseudo F and pseudo t^2 results from numerous methods are compared for consensus, the optimal number of clusters becomes distorted due to inconsistent information. This makes cluster analysis appear as more of an art form than a statistical tool. There are times when it is impossible to determine the optimal number of clusters.

This is not to say that cluster analysis does not have benefits. It provides the opportunity to look within groups to see the ordering of the objects. This is not possible with traditional statistical analyses. In ANOVA, MANOVA, and hypothesis testing, the groups are predetermined by the statistician. Inherent in these processes is a controlled allocation of group members. Whether these groups represent natural groupings is unresolved.

Another benefit of cluster analysis is prediction capability. Once ordering within clusters is established, cluster analysis may be used to predict placement of new objects into their proper clusters.

Most cluster analyses consist of calculating a Euclidean distance matrix from descriptive features. Creation of a distance matrix based on statistical test results may bring more statistical knowledge into the clustering process, but the issue of whether that helps form more stable clusters if unknown.

Biological Conclusions

Clustering of tumor suppressor genes and the *MYC* family based on the homology matrix is biologically significant. The clustering of oncogenes and non-regulatory genes based on number of introns, and to a lesser degree, results of the means of intronic log lengths combined with the number of introns, also indicate the occurrence of non-random processes. This finding supports the biological hypothesis that introns may have a specific purpose and they may group by degree of regulatory function.

The issue of the oncogenes consistently clustering closely with non-regulatory genes based on the homology matrix is an interesting result. It may indicate that there is not a substantial difference between these two groups. Hence, only the sequences of the tumor suppressor genes differ significantly from the other two groups based on sequence content. Another possibility is there is a closer relationship between oncogenes and non-regulatory genes than was previously thought. A possible hypothesis is that the intronic regions of the oncogenes somehow influence (maybe by transposons) the non-regulatory intronic regions. Why the intronic regions of these different gene groups are statistically similar is of great interest. Although statistical significance is not a guarantee of biological homology, it is usually considered to be an acceptable indicator of the existence of a relationship. A second possibility is that there are not yet enough genes in the sample for the distinctions between oncogenes and non-regulatory genes to be manifest.

The indication of Markov processes in some introns and not in others within the same gene is also of biological importance. It is possible that introns displaying Markov processes are more regulatory, while introns displaying randomness are less important to

80

gene function. Different introns may have displayed different levels of regulation at different evolutionary times. The results in this thesis also explain, to some degree, the inconsistency in the literature as to whether Markov processes occur in DNA sequences. We observed 67 percent Markov processes and 33 percent randomness. Depending upon the random sample taken, it is possible that either outcome could occur.

Areas of Future Interest

As more DNA data are sequenced, these results can be re-evaluated. It is possible that the clusters will separate more distinctly by regulatory group as the gene sample size increases. A look at clustering based only upon homology within the longer DNA intronic regions may also be of value. There are some intronic regions (e.g., intron 1) which are atypically long in many of the genes. Since they are noticeably different in length, they warrant further investigation. It will be interesting to see whether future laboratory studies and DNA analyses support our conclusions.

Use of Billingsley's homology statistic can be extended to three or more genes. This indicates the statistical possibility of multi-dimensional clustering. It might be possible, using the results of this statistic, to create multi-dimensional distance matrices. How these would compare with the two dimensional ones might lend more insight into the methodology behind present clustering processes.

The use of the homology statistic can be extended to higher order Markov processes. It was not possible in this thesis because there were many introns that were short, creating too many zero cells in higher order transition matrices. This approach could be investigated on the longer introns.

Results from a test of homology were used as a distance measure. I applied the same hypothesis test to all comparable intronic regions within a gene pair, regardless of whether they showed evidence of a zero-order or a first-order Markov process. For the comparison of two intronic regions where both regions are composed of independent sequences, one would normally do a Chi-square test of independence without the assumption of a Markov process. For the case where one intronic region is a zero-order Markov process and one intronic region is a first order Markov process, it might be appropriate to not compare the two sequences without knowledge of the underlying distribution. A methodology that incorporates the optimal hypothesis testing approach and provides insight into an underlying distribution for the test statistics would be advantageous. This methodology should result in the same degrees of freedom for all hypothesis test results so that the results can be used as a distance measure. In general, more methods that incorporate ordered data should be included into the clustering methodology and more statistical rigor should brought into the process.

Overall Conclusion

In conclusion, I have provided a methodology that allows the first systematic comparison of information within and between genes of various regulatory groupings. This offers an alternative to the current methodologies. The greatest strength of the methodology is the ability to apply it to the clustering of intronic regions, regions that have previously been extremely difficult to analyze and compare.

82

REFERENCES

Arques DG, Michel CJ. Periodicites in introns. *Nucleic Acids Research*, 15(18):7581-7592, 1987.

Arratia R, Morris P, Waterman MS. Stochastic Scrabble: Large Deviations for Sequences with Scores. *Journal of Applied Probability*, 25:106-119, 1988.

Altschul SF, Gish W, Miller W, Myers E, Lipman D. Basic Local Alignment Search Tool. *Journal of Molecular Biology*, 215:403-10, 1990.

Altschul SF, Boguski MS, Gish W, Wootton JC. Issues in Searching Molecular Sequence Databases. *Nature Genetics*, 6:119-129, 1994.

Authn H, Fieldes MA. A Set of BASIC Programs to Evaluate Relationships Among Protein Sequences by Optimum Alignment Matrix Analysis. *Computer Methods and Programs in Biomedicine*, 38(1):61-72, 1992.

Barrai I, Scapoli C, Barale R, Volina S. Oligonucleotide Correlations Between Infector and Host Genomes Hint at Evolutionary Relationships. *Nucleic Acids Research*, 18(10):3021-3025, 1990.

Baskin Y. Mapping the Cell's Nucleus. *Science*, 268:1564-1565, 1995.

Bharucha-Reid AT. *Elements of the Theory of Markov Processes and Their Applications.* McGraw-Hill Book Company, USA, 1960.

Bickam JW, Wood CC, Patton JC. Biogeographic Implications of Cytochrome b Sequences and Allozymes in Sockeye. *Journal of Heredity*, 86(2):140-144, 1995.

Billingsley P. Statistical Methods in Markov Chains. *Annals of Mathematical Statistics*, 32:12-40, 1961.

Boguski M. Computational Sequence Analysis Revisited: New Databases, Software Tools, and the Research Opportunies they Engender. *Journal of Lipid Research*, 33:957-972, 1992.

Bordonaro M, Nordstrom JL. Different Mechanisms are Responsible for the Low Accumulation of Transcripts from Intronless and 3'Splice Site Deleted Genes. *Biochemical and Biophysical Research Communications,* 203:128-32, 1994.

Bossu JP, Chartier FL, Vu-Dac N, Fruchart JC, Laine B. Transcription of the Human Apolipoprotein A-ll is Down-Regulated by the First Intron of its Gene. *Biochemical and Biophysical Research Communications,* 202:822-9, 1994.

Friedman JM, Dill FJ, Hayden MR, McGillivray BC. *Genetics*. Harwal Publishing Company, Pennsylvania, 1992.

Fuchs C. On the distribution of the nucleotides in seven completely sequenced DNAs. *Gene*, 10:371-373, 1980.

Garden PW. Markov analysis of viral DNA/RNA sequences. *Journal of Theoretical Biology*, 82:679-684, 1980.

Garrels JI, Franza BR, Chang C, Latter G. Quantitative exploration of the REF52 protein database: Cluster analysis reveals the major protein expression profiles in responses to growth regulation, serum stimulation, and viral transformation. *Electrophoresis*, 11(12):1114-1130, 1990.

Gatlin LL. *Information Theory and the Living System*. Columbia University Press, New York, 1972.

Gelfand M. Statistical analysis of pre-mRNA splicing sites. *Nucleic Acids Research*, 17(15):6369-6381, 1989.

Gelfand M, Kozhukhin CG, Pevzner PA. Extendable words in nucleotide sequences. *Computer Applications in the Biological Sciences*, 8(2):129-135, 1992.

Gentleman JF, Mullin RC. The distribution of the frequency of occurrence of nucleotide sequences, based on their overlap capability. *Biometrics*, 45(1):35-52, 1989.

Goodman SR. *Medical Cell Biology*. Philadelphia: J.B. Lippincott Company, 1994.

Hargrove JL, Schmidt FH. The role of mRNA and protein stability in gene expression. *The Federation of American Societies for Experimental Biology Journal*, 3:2360-2370, 1989.

Hasegawa M, Yano T. The genetic code and the entropy of protein. *Mathematical Biosciences*, 24:182-196, 1975.

Hartigan JA. *Clustering Algorithms*. John Wiley & Sons, Inc., New York, 1975.

Hawkins JD. A survey on intron and exon lengths. *Nucleic Acids Research*, 16(21):9893-9908, 1988.

Henikoff S, Henikoff JG. Protein family classification based on searching a database of blocks. *Genomics*, 19:97-107, 1994.

Hesketh R. *The Oncogene Facts Book*, Academic Press, Inc., London, 1994.

Hesketh R. *The Oncogene Handbook*, Academic Press, Inc., London, 1995.

Hickson RE. Self-splicing introns as a source for transposable genetic elements. *Journal of Theoretical Biology*, 141:1-10, 1989.

Hoel PG. A test for Markov chains. *Biometrika,* 41:430-433, 1954.

Johnson NL, Kotz S. *Continuous Univariate Distributions - 2.* Houghton Mifflin Company, Boston, pp. 94-148, 1970.

Johnston M. Sequencing the yeast genome. *University of Colorado Molecular Biology Program 1996 Annual Minocourse on DNA Sequence Analysis: Understanding and Using Computer Algorithms to Get the Most from your Sequences*, 1996.

Kaufman L, Rousseeuw PJ. *Finding Groups in Data: An Introduction to Cluster Analysis.* Wiley Series in Probability and Mathematical Statistics, USA, 1990.

Kelly C. A test of the Markovian model of DNA evolution. *Biometrics*, 50:653-664, 1994.

King A, Holmes B, Phillips I, Lapage SP. A taxonomic study of clinical isolates of Pseudomonas pickettii, 'P.thomasii', and 'Group Ivd' bacteria. *Journal of General Microbiology*, 114(1):137-147, 1979.

King RC, Stansfield WD. *A Dictionary of Genetics*, 4th Edition. Oxford University Press Inc., New York, 1990.

Kleffe J, Borodovsky M. First and second moments of counts of words in random texts generated by Markov chains. *Computer Applications in the Biosciences*, 8(5):433-441, 1992.

Kleffe J, Langbecker U. Exact Computation of Pattern Probabilities in Random Sequences Generated by Markov Chains. *Computer Applications in the Biosciences*, 6(4):347-353, 1990.

Konopka AK, Smythers GW. DISTAN - A program which detects significant distances between short oligonucleotides. *Computer Applications in the Biosciences*, 3(3):193-201, 1987.

Konopka, Smythers GW, Owens J, Maizel JV Jr. Distance analysis helps to establish characteristic motifs in intron sequences. *Gene Analysis Techniques*, 4(4):63-74, 1987.

Krogh A, Mian IS, Haussler D. A hidden Markov model that finds genes in *E. coli* DNA. *Nucleic Acids Research*, 22(22):4768-4778, 1994.

Kuiper FK, Fisher L. 391: *A Monte Carlo Comparison of Six Clustering Procedures*. Biometrics, 31:777-783, 1975.

Ladunga I. Phylogenetic continuum indicates "galaxies" in the protein universe. *Journal of Molecular Evolution*, 34(4):358-375, 1992.

Lake JA. Reconstructing evolutionary trees from DNA and protein sequences: Paralinear distances. *Proceedings of the National Academy of Sciences of the United States of America*, 91:1455-1459, 1994.

Legendre P, An introduction to classification and clustering. *Classification Society of North America Short Course CSNA-95, Session II:Cluster Analysis*, 1995.

Lewin B. *Genes V*. Oxford University Press Inc., New York, 1994.

Lewis EB, Knafels JD, Mathog DR, Celniker SE. Sequence analysis of the cis-regulatory regions of the bithorax complex of *Drosophila. Proceedings of the National Academy of Sciences of the United States of America*, 92:8403-8407, 1995.

Lipman DJ, Pearson WR. Rapid and sensitive protein similarity searches. *Science*, 227:1435-1441, 1985.

Lipshutz RJ, Teverner F, Hennessy K, Hartzell G, David R. DNA sequence confidence estimation. *Genomics*, 19:417-424, 1994.

Malkinson AM, You M. The intronic structure of cancer-related genes regulates susceptibility to cancer. *Molecular Carcinogenesis*, 10:61-65, 1994.

Mantegna RN, Buldyrev AL, Goldberger AL, Havlin S, Peng C-K, Simons M, and Stanley HE. Linguistic features of noncoding DNA sequences. *Physical Review Letters*, 73(23):3169-3172, 1994.

Meyer RJ. Mitochondrial DNAs and plasmids as taxonomic characteristics in trichoderma viride. *Applied and Environmental Microbiology*, 57(8):2269-2276, 1991.

Milligan GW. An introduction to classification and clustering. *Classification Society of North America Short Course CSNA-95, Session III: Cluster Validation and Comparison*, 1995.

Milligan GW. A Monte Carlo study of thirty internal criterion measures for cluster analysis. *Psychometrika*, 46(2):187-199, 1981.

Milligan GW. An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *Psychometrika*, 45:325-342, 1980.

Milligan GW, Cooper MC. An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50(2):159-179, 1985.

Mott RF, Kirkwood TBL, Curnow RN. A test for the statistical significance of DNA sequence similarities for application in databank searches. *Computer Applications in the Biosciences*, 5(2):123-131, 1989.

Mukwaya GM, Welch DF. Subgrouping of *Pseudomonas cepacia* by cellular fatty acid composition. *Journal of Clinical Microbiology*, 27(12):2640-2646, 1989.

Musser JM, Schlievert PM, Chow AW, Ewan P, Kreiswirth BN, Rosdahl VT, Naidu AS, Witte W, Selander RK. A single clone of *Staphylococcus aureus* causes the majority of cases of toxic shock syndrome. *Proceedings of the National Academy of Sciences of the United States of America*, 87(1):225-229, 1990.

Pearson ML, Soll D. The human genome project: A paradigm for information management in the life sciences. *The Federation of American Societies for Experimental Biology Journal*, 5:35-39, 1991.

Penotti FE. Human pre-mRNA splicing signals. *Journal of Theoretical Biology*, 150:385-420.

Phillips GJ, Arnold J, Ivarie R. Mono-through hexonucleotide composition of the *Escherichia coli* genome: a Markov chain analysis. *Nucleic Acids Research*, 15(6):2611-26, 1987.

Roberts L. Finding DNA sequencing errors. *Science*, 252:1255, 1991.

Romesburg CH. *Cluster Analysis for Researchers*. Kreiger Publishing Company, Florida, pp. 89-91, 1984.

Rowe GW, Szabo VL, Trainor LE. Cluster analysis in fenes in codon space. *Journal of Molecular Evolution*, 20(2):167-174, 1984.

Sakakibara Y, Brown M, Hughey R, Mian IS, Sjolander K, Underwood RC, Haussler D. Stochastic context-free grammers for tRNA modeling. *Nucleic Acids Research*, 22(23):5112-5120, 1994.

SAS Institute Inc. *SAS Guide to Macro Processing*, Version 6.0 Edition, SAS Institute Inc., Cary, North Carolina, 1989.

SAS Institute Inc. *SAS/Stat User's Guide, Release 6.03 Edition*, SAS Institute Inc., Cary, North Carolina, pp. 283-357, 1989.

Schiavi SC, Belasco JG, Greenberg ME. Regulation of proto-oncogene mRNA stability. *Biochimica et Biophysica Acta*, 1114:95-106, 1992.

Smith EL, Hill RL, Lehman RI, Lefkowitz RJ, Handler P, White A. *Principles of Biochemistry: General Aspects*, 7th Edition. MacGraw-Hill, Inc., USA, pp. 179-237, 387-478.

Snyder EE, Stormo GD. Identification of coding regions in genomic DNA sequences: An application of dynamic programming and neural networks. *Nucleic Acids Research*, 21:607-613, 1993.

Snyder EE. Personal conversation, 1995.

Telliez J-B, Plumb M, Balmain A, Bailleul B. Regulatory elements in the first intron of the mouse Ha-*ras* gene. *Molecular Carcinogenesis*, 12:137-145, 1995.

Tong H. Determination of the order of a Markov chain by Akaike's information criterion. *Journal of Applied Probability*, 12:488-497, 1975.

van der Mei HC, de Soet JJ, de Graaff J, Rouxhet PG, Busscher HJ.Comparison of the physicochemical surface properties of S*treptococcus rattus* with those of other mutant streptococcal species. *Caries Research*, 25(6):415-423, 1991.

Watson JD, Gilman M, Witkowski J, Zoller M. *Recombinant DNA*, 2nd Edition, W.H. Freeman and Company, 66-68, 1992.

Weir BS. Analysis of DNA sequences. *Statistical Methods in Medical Research*, 2:225-239, 1993.

Wills C. *Exons, Introns and Talking Genes*. New York: Basic Books, 1991.

You M, Wang Y, Stoner GD, et al. Parental bias of K-*ras* oncogenes detected in lung tumors from mouse hybrids. *Proceedings of the National Academy of Sciences of the United States of America*, 89:5804-5808, 1992.

Zerbe GO, Murphy JR. On multiple comparisons in the randomization analysis of growth and response curves. *Biometrics*, 42:795-804, 1986.

APPENDIX A

MATERIAL TO SUPPLEMENT CHAPTER IV
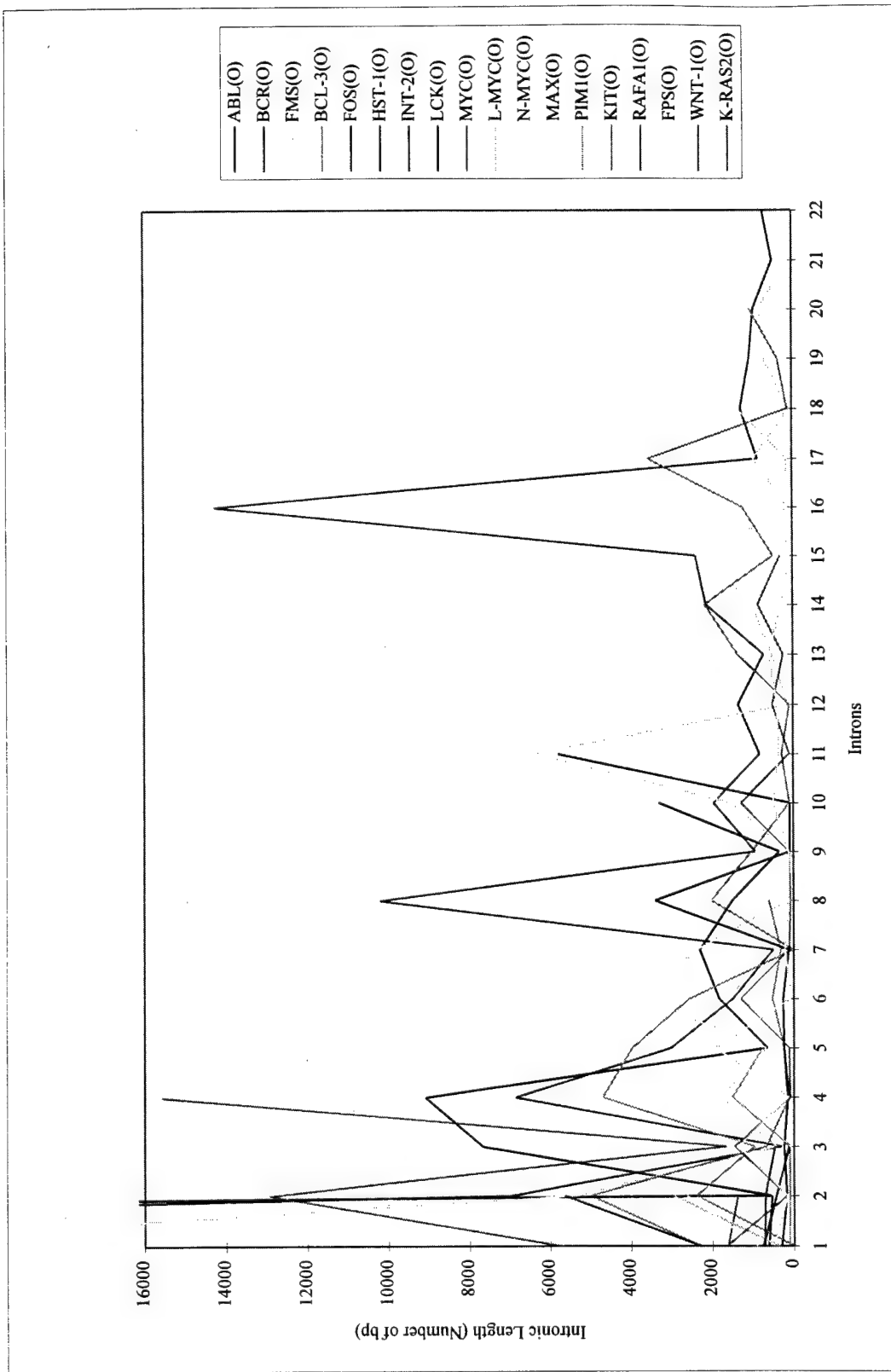
Figure A.1: Graph of Intronic Lengths for Oncogenes Truncated at 16,000 bp
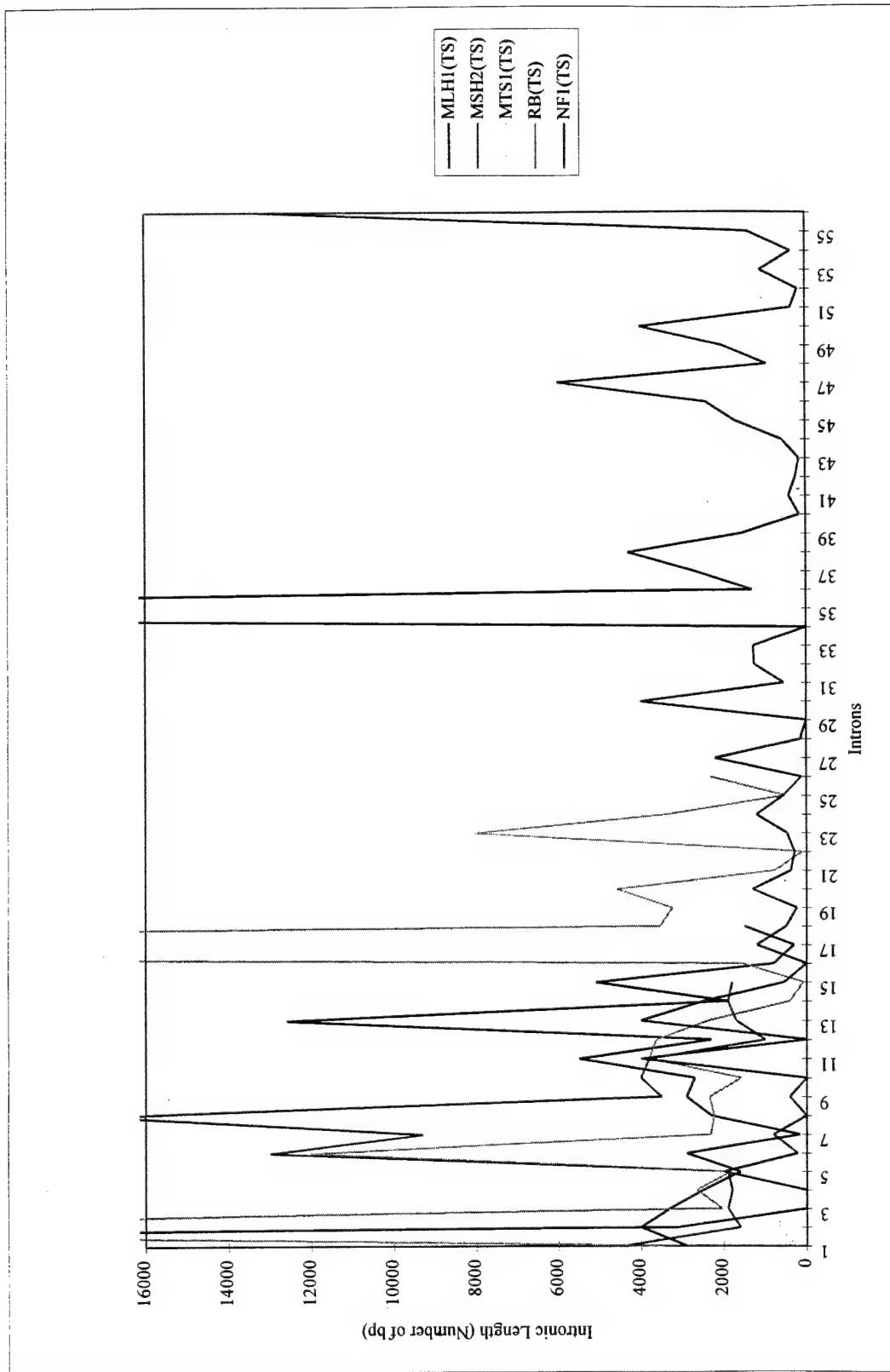
Figure A.2: Graph of Intronic Lengths for Tumor Suppressor Genes Truncated at 16,000 bp
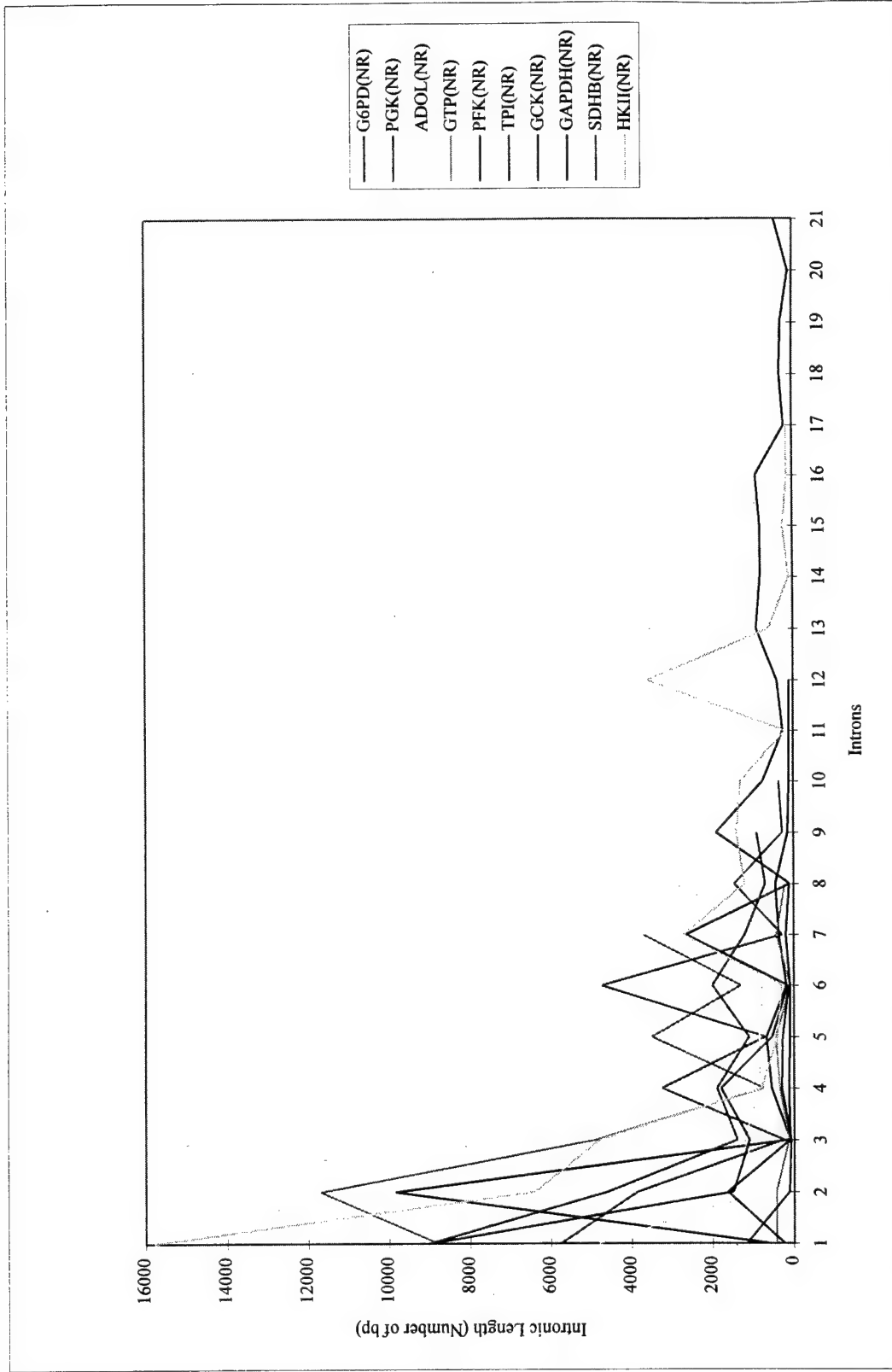
Figure A.3: Graph of Intronic Lengths for Non-Regulatory Genes Truncated at 16,000 bp
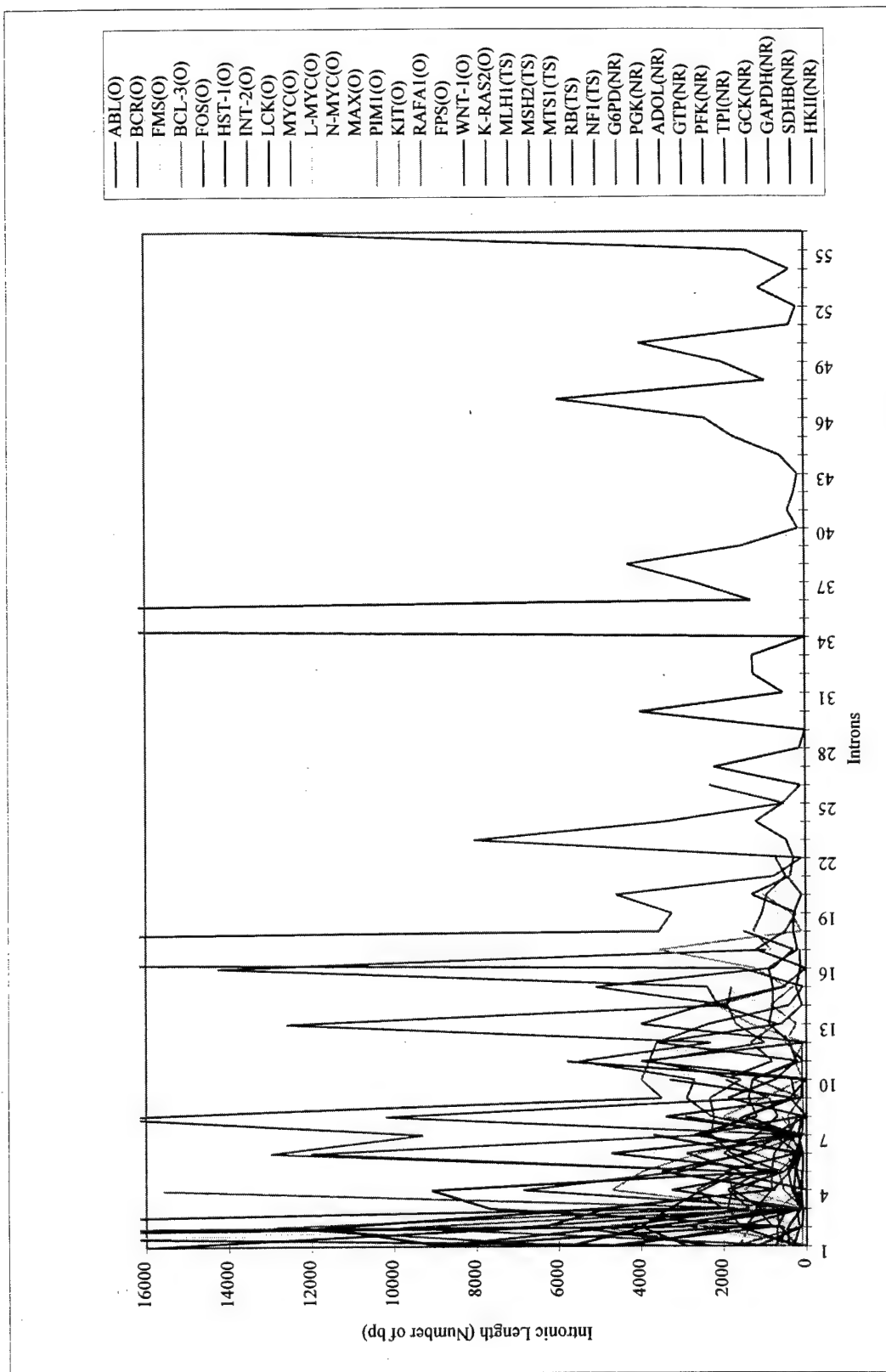
Figure A.4: Graph of Intronic Lengths for All Genes in Data Set Truncated at 16,000 bp

93
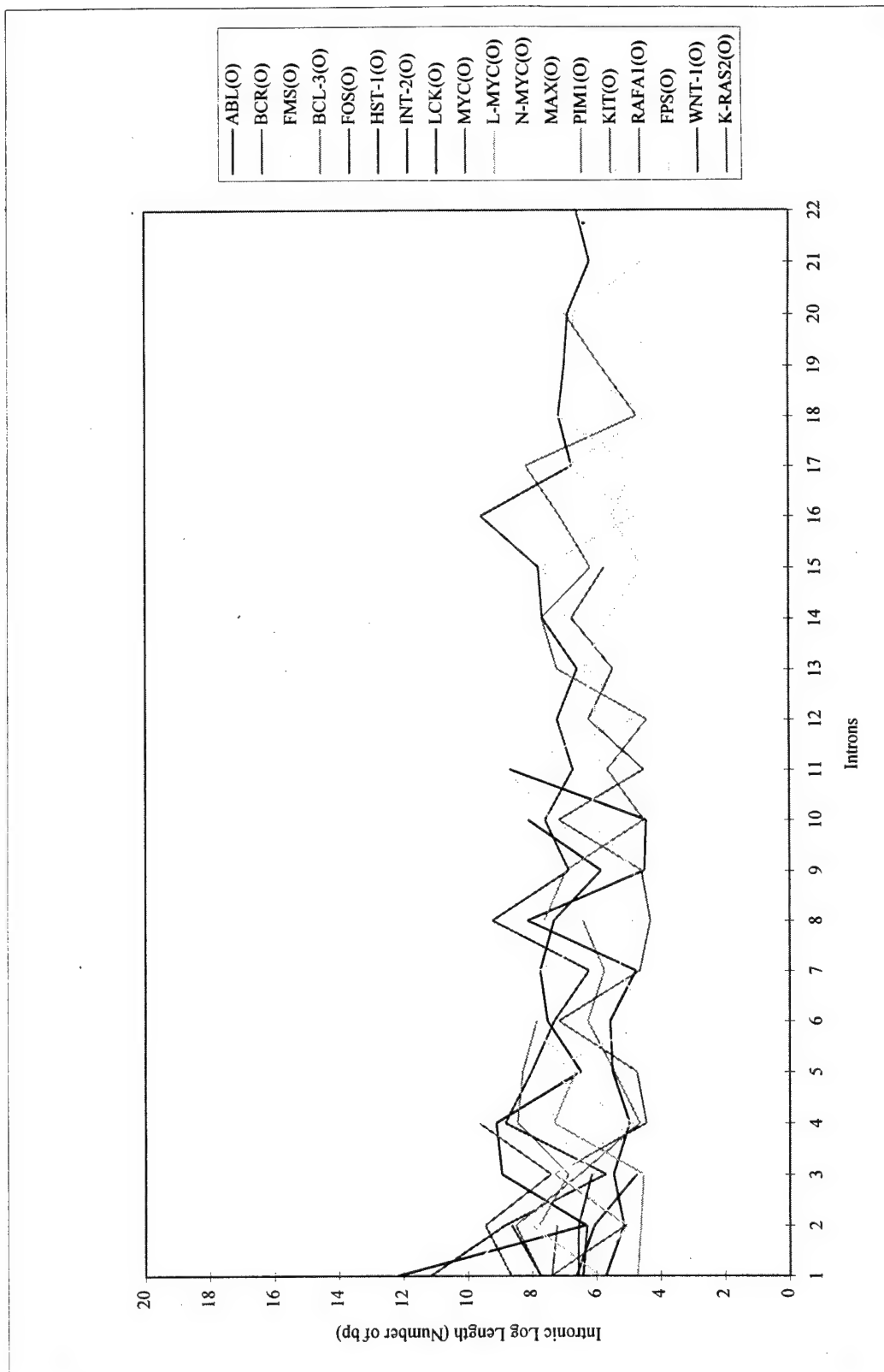
Figure A.5: Graph of Intronic Log Lengths for Oncogenes

94
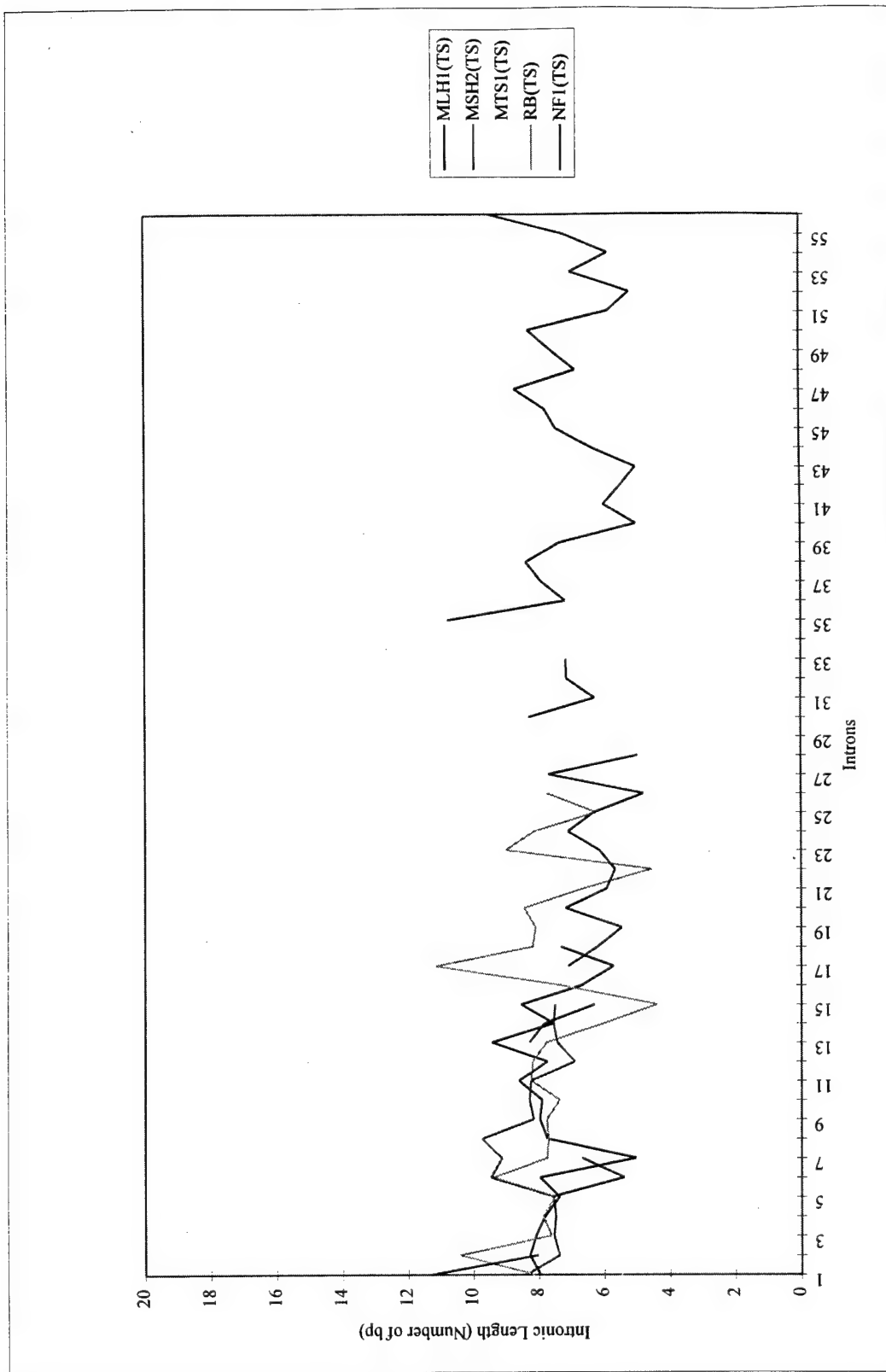
Figure A.6: Graph of Intronic Log Lengths for Tumor Suppressor Genes
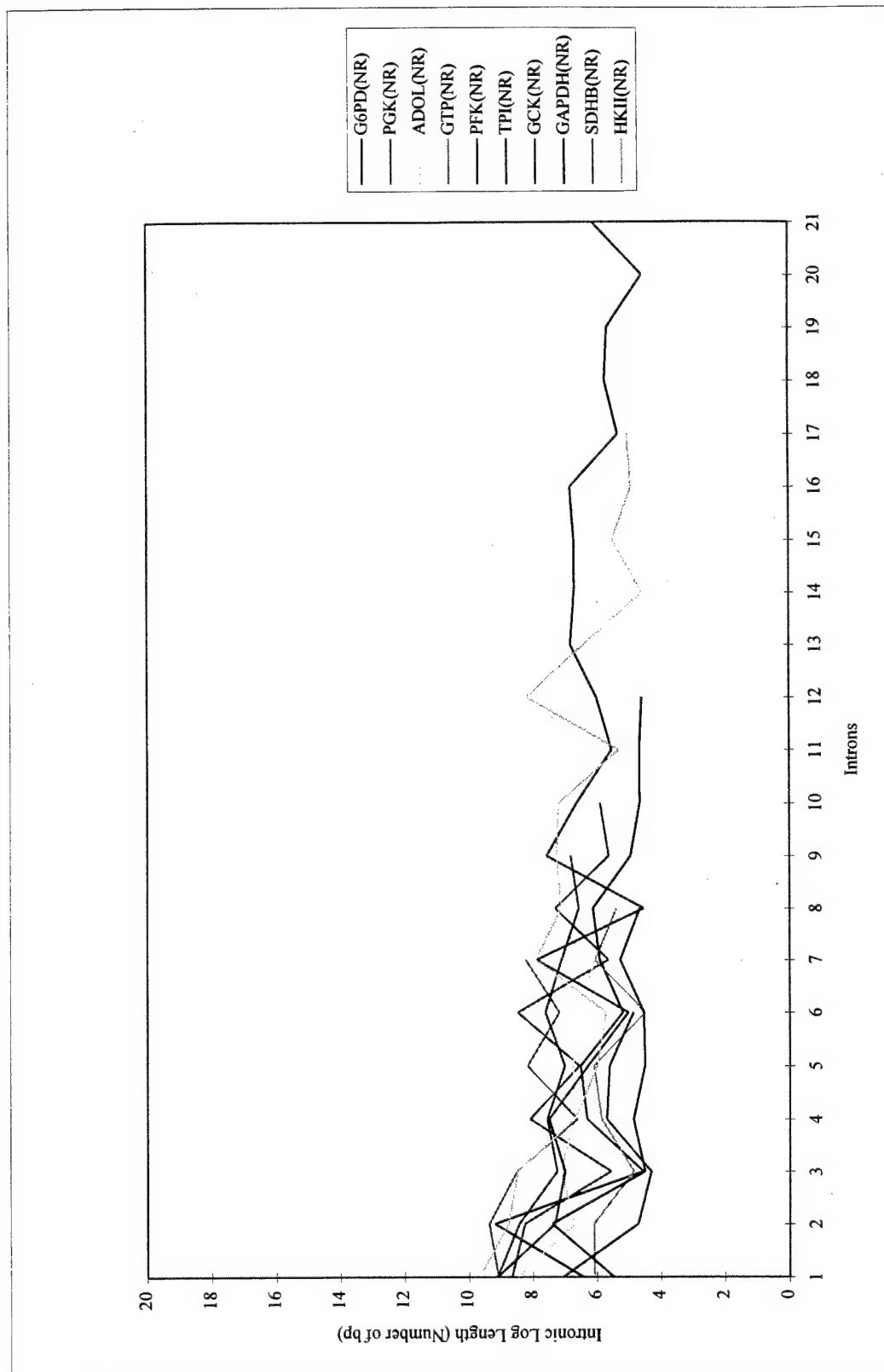
Figure A.7: Graph of Intronic Log Lengths for Non-Regulatory Genes
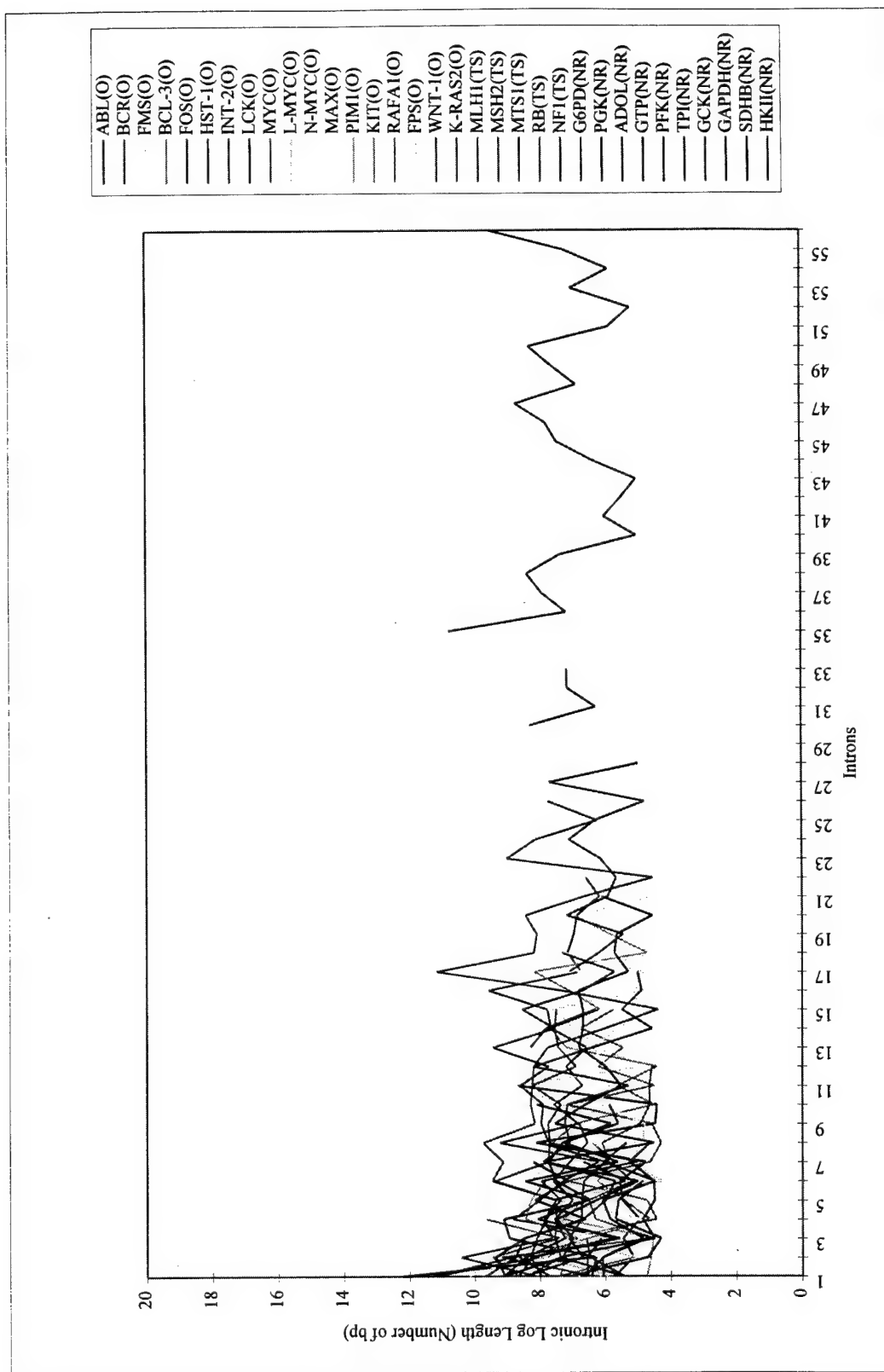
Figure A.8: Graph of Intronic Log Lengths for All Genes in Data Set

Legend:
ABL(O)
BCR(O)
FMS(O)
BCL-3(O)
FOS(O)
HST-1(O)
INT-2(O)
LCK(O)
MYC(O)
L-MYC(O)
N-MYC(O)
MAX(O)
PIM1(O)
KIT(O)
RAFA1(O)
FPS(O)
WNT-1(O)
K-RAS2(O)
MLH1(TS)
MSH2(TS)
MTS1(TS)
RB(TS)
NF1(TS)
G6PD(NR)
PGK(NR)
ADOL(NR)
GTP(NR)
PFK(NR)
TPI(NR)
GCK(NR)
GAPDH(NR)
SDHB(NR)
HKII(NR)

Y-axis: Intronic Log Length (Number of bp), values 0, 2, 4, 6, 8, 10, 12, 14, 16, 18, 20

X-axis: Introns, values 1, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, 25, 27, 29, 31, 33, 35, 37, 39, 41, 43, 45, 47, 49, 51, 53, 55

Table A.1: Means and Standard Deviations of Intronic Lengths and Intronic Log Lengths

Gene	Mean of Intronic Lengths in a Gene	Standard Deviation of Intronic Lengths in a Gene	Mean of Intronic Log Lengths in a Gene	Standard Deviation of Intronic Log Lengths in a Gene
ABL(O)	22724.30	62361.11	7.96	1.84
BCR(O)	5943.00	15089.64	7.53	1.30
FMS(O)	2622.10	5632.93	6.68	1.57
BCL-3(O)	1218.63	1676.15	6.41	1.24
FOS(O)	432.67	319.50	5.81	0.97
HST-1(O)	577.50	55.86	6.36	0.10
INT-2(O)	3968.50	2375.17	8.19	0.64
LCK(O)	986.09	1867.00	5.71	1.40
MYC(O)	1500.00	175.36	7.31	0.12
L-MYC(O)	1667.50	1843.43	6.95	1.48
N-MYC(O)	1765.00	1231.78	7.34	0.76
MAX(O)	483.00	NA	6.18	NA
PIM1(O)	514.40	622.44	5.57	1.31
KIT(O)	1574.17	1408.35	6.79	1.30
RAFA1(O)	554.00	583.70	5.70	1.19
FPS(O)	473.56	531.84	5.68	0.97
WNT-1(O)	625.67	141.85	6.42	0.25
K-RAS2(O)	8979.75	6414.46	8.80	1.02
MLH1(TS)	3069.67	2774.06	7.65	1.02
MSH2(TS)	4573.33	4755.22	8.05	0.84
MTS1(TS)	536.67	266.58	6.17	0.63
RB(TS)	6644.96	14556.52	7.75	1.46
NF1(TS)	4285.98	13155.08	6.99	1.41
G6PD(NR)	1102.50	2765.38	5.75	1.34
PGK(NR)	2092.10	2110.98	7.00	1.29
ADOL(NR)	1581.75	1516.88	7.05	0.80
GTP(NR)	315.00	148.81	5.61	0.63
PFK(NR)	1182.86	1915.21	6.41	1.14
TPI(NR)	341.17	413.58	5.37	0.99
GCK(NR)	2534.67	2694.33	7.49	0.82
GAPDH(NR)	321.50	533.19	5.15	0.98
SDHB(NR)	4950.00	3979.43	8.16	0.98
HKII(NR)	2360.06	3918.63	6.74	1.50

APPENDIX B

MATERIAL TO SUPPLEMENT CHAPTER V

Table B.1: Markov Results Superimposed over Lengths of DNA Sequence Data Available For Introns in Each Gene

Gene	# of Introns In Gene	Introns															
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
ABL(O)	10	103531	563	7666	9093	646	1830	2297	1500	342	3306						
BCR(O)	22	71559	6969	300	6867	3016	1488	500	10202	934	1957	818	1344	717	2127	2385	14263
FMS(O)	21	1735	5349	430	1803	676	3780	2370	153	118	1542	6355	127	517	999	2107	1119
BCL-3(O)	8	1355	1153	676	105	230	522	310	606								
FOS(O)	3	753	431	114													
HST-1(O)	2	617	538														
INT-2(O)	2	2289	5648														
LCK(O)	11	304	172	236	143	241	261	117	510	89	84	443					
MYC(O)	2	1624	1376														
L-MYC(O)	2	364	2971														
N-MYC(O)	2	894	2636														
MAX(O)	1	483															
PIM1(O)	5	113	101	93	1504	761											
KIT(O)	20	226	194	160	408	273	557	260	414	569	91	280	83	549	363	346	427
RAFA1(O)	15	1652	162	1458	85	114	1290	103	74	98	1281	91	505	236	853	308	
FPS(O)	18	492	152	1373	129	1931	73	203	129	121	495	361	381	594	274	96	239
WNT-1(O)	3	713	702	462													
K-RAS2(O)	4	246	441	551	1181												
MLH1(TS)	18	287	267	265	200	262	358	154	494	328	349	256	245	335	325	178	172
MSH2(TS)	15	282	182	245	181	187	240	162	397	651	792	775	418	551	140	442	
MTS1(TS)	3	231	658	721													
RB(TS)	26	530	550	448	548	429	522	513	432	438	581	606	524	567	403	80	550
NF1(TS)	56	47	135	374	373	427	220	249	235	356	108	1438	139	400	560	194	194
G6PD(NR)	12	625	9856	95	549	671	177	365	447	139	104	105	97				
PGK(NR)	10	451	260	211	355	200	227	178	257	123	249						
ADOL(NR)	8	4811	809	1231	836	841	278	411	2937								
GTP(NR)	8	435	440	128	350	440	89	424	214								
PFK(NR)	21	87	89	110	192	178	150	1065	92	172	312	188	77	94	44	50	34
TPI(NR)	6	1165	111	74	297	272	128										
GCK(NR)	9	2009	367	310	398	109	253	356	277	330							
GAPDH(NR)	8	240	1634	90	129	90	92	193	104								
SDHB(NR)	7	631	209	618	483	378	294	334									
HKII(NR)	17	107	187	203	217	199	136	205	203	183	207	165	102	238	96	242	133

This table superimposes results from a test of a zero-order Markov process versus a first-order Markov process over the sizes of the intronic regions in an attempt to determine whether only short intronic regions show evidence of a zero-order Markov process. Gray cells indicate that the intronic region showed statistical evidence of a first-order Markov process at the $\alpha = .01$ level while white cells indicate that intronic regions failed to reject the null hypothesis of a zero-order Markov process.

Table B.1 (Continued): Markov Results Superimposed over Lengths of DNA Sequence Data Available for Introns in Each Gene

Gene	Introns																			
	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36
ABL(O)	840	1263	1050	946	478	718														
BCR(O)	945	81	690	806	97															
FMS(O)																				
BCL-3(O)																				
FOS(O)																				
HST-1(O)																				
INT-2(O)																				
LCK(O)																				
MYC(O)																				
L-MYC(O)																				
N-MYC(O)																				
MAX(O)																				
PIM1(O)																				
KIT(O)	143	111	350	237																
RAFA1(O)	124	1357																		
FPS(O)																				
WNT-1(O)																				
K-RAS2(O)	219	245																		
MLH1(TS)																				
MSH2(TS)																				
MTS1(TS)																				
RB(TS)	488	542	439	582	561	93	634	594	511	604										
NF1(TS)	183	490	231	109	369	285	461	783	344	120	132	145	318	1269	532	1039	227	211	37027	1256
G6PD(NR)																				
PGK(NR)																				
ADOL(NR)																				
GTP(NR)																				
PFK(NR)	99	55	89	77	229															
TPI(NR)																				
GLUC(NR)																				
GAPDH(NR)																				
SDHB(NR)																				
HKII(NR)	150																			

Table B.1 (Continued): Markov Results Superimposed over Lengths of DNA Sequence Data Available For Introns in Each Gene

Gene	Introns																			
	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56
ABL(O)																				
BCR(O)																				
FMS(O)																				
BCL-3(O)																				
FOS(O)																				
HST-1(O)																				
INT-2(O)																				
LCK(O)																				
MYC(O)																				
L-MYC(O)																				
N-MYC(O)																				
MAX(O)																				
PIM1(O)																				
KIT(O)																				
RAFA1(O)																				
FPS(O)																				
WNT-1(O)																				
K-RAS2(O)																				
MLH1(TS)																				
MSH2(TS)																				
MTSI(TS)																				
RB(TS)	2706	4324	1408	162	453	237	144	569	1699	2367	5986	931	1939	4055	377	178	1110	346	1471	13296
NF1(TS)																				
G6PD(NR)																				
PGK(NR)																				
ADOL(NR)																				
GTP(NR)																				
PFK(NR)																				
TPI(NR)																				
GLUC(NR)																				
GAPDH(NR)																				
SDHB(NR)																				
HKII(NR)																				

APPENDIX C

MATERIAL TO SUPPLEMENT CHAPTER VI

Table C.1: ABL(O) Gene (10 Introns) Versus Contrast Genes (Gray Cells Reject Ho at α = 0.01 Level))

Contrast Gene	# Introns in Contrast Gene	# Introns Compared	Intron 1	Intron 2	Intron 3	Intron 4	Intron 5	Intron 6	Intron 7	Intron 8	Intron 9	Intron 10	Mean	SD
BCR(O)	22	10	1874.80	121.65	12.64	129.06	21.68	38.50	38.21	234.29	17.72	32.20	252.08	574.53
FMS(O)	21	10	32.56	72.14	29.68	120.87	36.06	75.27	56.30	60.60	15.36	37.77	53.66	30.60
BCL-3(O)	8	8	214.60	38.50	261.01	31.47	37.43	146.11	60.11	251.40			130.08	100.61
FOS(O)	3	3	222.63	30.15	13.44								88.74	116.25
HST-1(O)	2	2	379.40	94.15									236.78	201.70
INT-2(O)	2	2	847.61	184.91									516.26	468.60
LCK(O)	11	10	61.64	47.79	62.24	47.63	51.60	47.05	32.09	52.07	12.35	8.98	42.34	18.70
MYC(O)	2	2	431.10	31.57									231.33	282.51
L-MYC(O)	2	2	269.55	31.83									150.69	168.10
N-MYC(O)	2	2	172.87	31.47									102.17	99.98
MAX(O)	1	1	94.49										94.49	NA
PIM1(O)	5	5	84.47	66.66	46.34	49.14	29.33						55.19	21.05
KIT(O)	20	10	43.83	11.76	22.73	52.26	52.66	25.07	11.25	34.46	34.69	17.93	30.66	15.44
RAFA1(O)	15	10	160.74	24.67	25.69	25.42	20.91	109.17	9.97	38.01	14.78	45.79	47.51	48.75
FPS(O)	18	10	197.03	60.92	46.70	21.61	30.60	25.78	39.32	57.35	38.27	39.88	55.75	51.20
WNT-1(O)	3	3	157.15	94.77	91.46								114.46	37.01
K-RAS2(O)	4	4	122.13	33.07	67.86	140.82							90.97	49.47
MLH1(TS)	18	10	60.62	6.09	21.83	48.87	41.68	31.39	25.89	33.16	20.01	19.55	30.91	15.98
MSH2(TS)	15	10	69.66	37.15	45.15	96.60	46.98	87.53	36.37	42.22	33.72	96.26	59.16	25.77
MTS1(TS)	3	3	74.75	90.99	26.85								64.20	33.34
RB(TS)	26	10	66.01	12.64	61.63	119.57	68.41	90.73	99.96	35.04	49.15	139.72	74.29	38.73
NF1(TS)	56	10	24.74	19.93	61.52	21.15	76.12	48.06	35.54	42.57	39.79	56.09	42.55	18.39
G6PD(NR)	12	10	86.96	80.03	38.39	79.58	46.95	46.05	26.70	101.33	16.61	19.65	54.23	30.44
PGK(NR)	10	10	25.64	9.07	29.02	18.50	34.73	21.43	31.14	17.91	22.43	27.47	23.73	7.49
ADOL(NR)	8	8	69.54	34.27	57.03	39.79	49.89	12.69	35.10	70.33			46.08	19.59
GTP(NR)	8	8	49.50	15.81	15.65	16.39	25.30	18.31	15.96	38.71			24.46	12.86
PFK(NR)	21	10	34.30	26.72	28.20	69.98	23.18	37.41	140.16	50.24	9.03	43.79	46.30	36.89
TPI(NR)	6	6	372.93	29.09	15.94	28.57	22.41	22.61					81.92	142.65
GCK(NR)	9	9	331.26	65.76	82.12	37.45	39.84	32.52	27.55	124.10	33.15		85.97	97.20
GAPDH(NR)	8	8	237.46	108.72	17.37	28.63	10.26	21.47	27.44	27.10			59.81	78.17
SDHB(NR)	7	7	34.08	9.43	38.76	18.87	7.66	9.25	20.40				19.78	12.46
HKII(NR)	17	10	40.31	12.88	16.58	26.10	30.15	41.97	17.45	36.24	8.98	20.17	25.08	11.70

Tables C.1 -C.33 depict test of homology results for intronic regions in each of the 33 genes compared against all other corresponding intronic regions from all genes in the data set. The number of introns compared is based on the gene with the fewest intronic regions. Gray cells indicate that the two intronic regions compared are not homologous (at the α = 0.01 level) while white cells indicate that the two regions fail to reject the null hypothesis of statistical homology.

104

Table C.2: BCR(O) Gene (22 Introns) Versus Contrast Genes (Gray Cells Reject Ho at α = 0.01 Level)

Contrast Gene	# Introns in Contrast Gene	# Introns Compared	Intron 1	Intron 2	Intron 3	Intron 4	Intron 5	Intron 6	Intron 7	Intron 8	Intron 9	Intron 10	Intron 11
ABL(O)	10	10	1874.8	121.65	12.64	129.06	21.68	38.5	38.21	234.29	17.72	32.2	
FMS(O)	21	21	103.38	58.31	8.50	76.59	22.28	72.67	41.23	19.65	13.20	27.64	40.98
BCL-3(O)	8	8	95.25	55.79	67.39	17.32	56.25	114.38	21.00	207.25			
FOS(O)	3	3	133.19	11.38	12.60								
HST-1(O)	2	2	269.46	19.55									
INT-2(O)	2	2	512.28	131.21									
LCK(O)	11	11	39.57	15.04	29.69	25.77	105.64	29.85	21.31	23.35	14.89	4.92	52.97
MYC(O)	2	2	288.83	88.54									
L-MYC(O)	2	2	198.14	106.87									
N-MYC(O)	2	2	75.70	100.87									
MAX(O)	1	1	31.10										
PIM1(O)	5	5	61.08	85.61	26.68	68.32	54.33						
KIT(O)	20	20	31.78	44.59	26.77	108.42	100.57	44.54	19.61	187.17	98.28	26.80	67.27
RAFA1(O)	15	15	63.84	15.18	20.88	18.36	14.84	201.33	9.07	22.44	6.78	66.98	19.65
FPS(O)	18	18	100.64	32.95	7.11	11.78	52.30	16.49	15.49	36.31	35.05	24.45	20.46
WNT-1(O)	3	3	84.48	90.38	28.86								
K-RAS2(O)	4	4	101.01	117.92	41.40	296.82							
MLH1(TS)	18	18	61.40	59.99	25.73	80.51	83.17	63.41	44.69	140.67	53.63	33.27	86.66
MSH2(TS)	15	15	48.91	144.97	41.82	148.74	91.31	133.13	60.03	231.16	100.21	141.10	135.46
MTS1(TS)	3	3	40.37	26.35	7.91								
RB(TS)	26	22	74.50	187.22	44.82	218.95	145.29	174.65	94.43	199.95	134.15	179.95	129.08
NF1(TS)	56	22	38.60	49.84	40.08	55.58	170.32	84.30	62.32	181.95	110.47	71.78	197.99
G6PD(NR)	12	12	40.84	52.48	21.53	43.71	44.15	25.29	11.56	22.28	14.55	12.79	14.04
PGK(NR)	10	10	23.54	64.43	25.80	48.28	54.22	40.05	42.34	54.57	45.31	25.38	
ADOL(NR)	8	8	226.34	65.30	30.21	109.66	106.37	27.06	42.34	458.35			
GTP(NR)	8	8	15.73	47.53	10.36	18.77	72.07	8.54	31.60	46.03			
PFK(NR)	21	21	17.22	7.92	18.31	39.86	13.59	24.02	40.08	22.11	6.24	23.34	10.21
TPI(NR)	6	6	210.59	11.02	17.04	18.32	14.27	12.08					
GCK(NR)	9	9	121.53	15.41	27.62	25.77	27.32	15.56	20.81	76.34	30.71		
GAPDH(NR)	8	8	174.22	79.07	5.74	28.36	7.71	15.47	16.96	16.90			
SDHB(NR)	7	7	19.24	50.95	11.26	47.11	12.04	15.02	23.39				
HKII(NR)	17	17	27.38	24.68	12.57	21.87	30.71	28.94	13.42	19.70	6.34	23.22	11.23

Table C.2 (Continued): BCR(O) Gene (22 Introns) Versus Contrast Genes (Gray Cells Reject Ho at $\alpha = 0.01$ level))

Contrast Gene	Intron 12	Intron 13	Intron 14	Intron 15	Intron 16	Intron 17	Intron 18	Intron 19	Intron 20	Intron 21	Intron 22	Mean	SD
ABL(O)												252.08	574.53
FMS(O)	9.64	35.02	12.74	30.35	18.16	8.80	15.38	27.26	58.06	22.47		34.40	25.82
BCL-3(O)												79.33	61.32
FOS(O)												52.39	69.98
HST-1(O)												144.50	176.72
INT-2(O)												321.75	269.45
LCK(O)												33.00	27.32
MYC(O)												188.69	141.63
L-MYC(O)												152.51	64.54
N-MYC(O)												88.29	17.80
MAX(O)												31.10	NA
PIM1(O)												59.20	21.60
KIT(O)	32.39	70.54	64.80	32.94	125.08	27.12	22.99	67.66	35.17			61.72	43.36
RAFA1(O)	43.31	34.56	67.63	85.29								46.01	49.64
FPS(O)	7.06	22.66	18.23	29.55	23.24	34.10	26.63					28.58	21.27
WNT-1(O)												67.91	33.94
K-RAS2(O)												139.29	110.03
MLH1(TS)	46.90	38.50	30.02	17.93	11.42	34.11	20.23					51.79	31.67
MSH2(TS)	121.61	76.92	26.75	75.83								105.20	53.36
MTS1(TS)												24.88	16.28
RB(TS)	203.48	93.75	44.25	74.44	137.83	66.11	163.11	92.49	146.12	189.88	37.38	128.72	57.16
NF1(TS)	81.96	54.41	133.99	77.08	53.40	54.77	110.27	54.11	68.08	102.25	82.21	87.99	45.96
G6PD(NR)	16.72											26.66	14.58
PGK(NR)												42.39	13.91
ADOL(NR)												133.20	146.59
GTP(NR)												31.33	22.35
PFK(NR)	9.09	24.77	8.86	21.42	15.10	39.91	9.71	7.77	6.72	26.63		18.71	11.05
TPI(NR)												47.22	80.08
GCK(NR)												40.12	35.60
GAPDH(NR)												43.05	57.92
SDHB(NR)												25.57	16.60
HKII(NR)	13.07	29.16	13.31	7.13	6.39	23.63						18.40	8.46

Table C.3: FMS(O) Gene (21 Introns) Versus Contrast Genes (Gray Cells Reject Ho at α= 0.01 Level)

Contrast Gene	# Introns in Contrast Gene	# Introns Compared	Intron 1	Intron 2	Intron 3	Intron 4	Intron 5	Intron 6	Intron 7	Intron 8	Intron 9	Intron 10	Intron 11
ABL(O)	10	10	32.56	72.14	29.68	120.87	36.06	75.27	56.3	60.6	15.36	37.77	
BCR(O)	22	21	103.38	58.31	8.50	76.59	22.28	72.67	41.23	19.65	13.20	27.64	40.98
BCL-3(O)	8	8	97.08	40.18	70.34	36.17	61.34	264.35	83.89	81.25			
FOS(O)	3	3	122.27	4.51	22.29								
HST-1(O)	2	2	224.35	35.34									
INT-2(O)	2	2	311.19	217.32									
LCK(O)	11	11	39.62	26.81	30.17	32.63	93.63	67.97	40.60	23.30	7.78	7.15	26.67
MYC(O)	2	2	165.74	47.03									
L-MYC(O)	2	2	184.71	39.05									
N-MYC(O)	2	2	120.67	43.29									
MAX(O)	1	1	71.42										
PIM1(O)	5	5	83.06	96.68	27.49	80.43	29.40						
KIT(O)	20	20	42.14	32.43	35.02	93.57	69.91	72.35	13.05	62.57	46.69	30.67	48.66
RAFA1(O)	15	15	89.20	13.50	36.42	14.77	10.57	219.70	14.09	25.60	9.16	89.00	18.83
FPS(O)	18	18	156.56	37.16	9.50	15.23	17.78	34.13	56.03	28.61	18.38	17.97	29.03
WNT-1(O)	3	3	108.20	116.51	33.71								
K-RAS2(O)	4	4	104.41	91.08	79.54	213.12							
MLH1(TS)	18	18	57.32	35.04	43.05	73.40	53.51	57.67	24.42	49.73	32.10	36.06	89.11
MSH2(TS)	15	15	67.34	111.08	64.01	112.30	71.93	120.70	38.89	78.17	47.08	148.17	147.22
MTS1(TS)	3	3	54.48	45.53	14.27								
RB(TS)	26	21	58.01	127.35	71.11	178.72	80.00	154.20	67.59	81.17	61.21	194.19	137.95
NF1(TS)	56	21	34.83	39.18	67.66	64.68	102.82	65.85	32.35	65.34	50.75	77.99	277.17
G6PD(NR)	12	12	63.92	60.19	19.08	56.69	44.39	53.97	29.34	21.01	7.68	13.12	25.04
PGK(NR)	10	10	37.75	44.80	36.05	51.17	28.66	30.96	28.08	40.49	29.59	40.65	
ADOL(NR)	8	8	9.85	35.51	40.14	87.12	43.03	25.00	18.98	62.01			
GTP(NR)	8	8	38.05	32.95	16.48	21.75	51.05	21.12	14.80	38.41			
PFK(NR)	21	21	33.51	13.08	13.44	38.14	20.39	38.30	203.66	15.95	8.44	25.18	26.82
TPI(NR)	6	6	195.41	9.67	20.92	15.85	5.59	19.11					
GCK(NR)	9	9	167.64	29.55	32.60	21.75	28.35	41.74	37.15	11.39	24.71		
GAPDH(NR)	8	8	166.89	113.56	7.00	24.72	15.56	34.22	32.60	13.18			
SDHB(NR)	7	7	31.42	33.08	24.05	41.03	13.29	31.38	19.92				
HKII(NR)	17	17	43.07	15.55	7.32	17.18	23.83	36.58	41.76	14.68	7.56	34.71	23.51

Table C.3 (Continued): FMS(O) Gene (21 Introns) Versus Contrast Genes (Gray Cells Reject Ho at α= 0.01 Level)

Contrast Gene	Intron 12	Intron 13	Intron 14	Intron 15	Intron 16	Intron 17	Intron 18	Intron 19	Intron 20	Intron 21	Mean	SD
ABL(O)											53.66	30.60
BCR(O)	9.64	35.02	12.74	30.35	18.16	8.80	15.38	27.26	58.06	22.47	34.40	25.82
BCL-3(O)											91.83	72.82
FOS(O)											49.69	63.48
HST-1(O)											129.84	133.65
INT-2(O)											264.25	66.37
LCK(O)											36.03	25.27
MYC(O)											106.39	83.94
L-MYC(O)											111.88	102.99
N-MYC(O)											81.98	54.72
MAX(O)											71.42	NA
PIM1(O)											63.41	32.52
KIT(O)	25.66	84.82	56.60	37.76	56.49	29.92	23.62	58.48	7.08		46.38	23.04
RAFA1(O)	11.18	24.53	46.46	83.04							47.07	55.95
FPS(O)	5.95	40.93	23.97	34.01	7.88	37.00	14.97				32.50	33.64
WNT-1(O)											86.14	45.60
K-RAS2(O)											122.04	61.57
MLH1(TS)	29.04	38.16	26.62	25.28	18.35	33.49	21.50				41.33	18.96
MSH2(TS)	42.27	88.59	25.90	80.00							82.91	38.16
MTS1(TS)											38.09	21.11
RB(TS)	75.51	93.53	47.72	91.05	51.94	76.80	51.12	60.08	39.98	77.12	89.35	43.66
NF1(TS)	53.39	70.07	118.23	85.73	39.36	56.38	34.03	37.09	25.56	31.69	68.10	53.77
G6PD(NR)	11.16										33.80	20.77
PGK(NR)											36.82	7.67
ADOL(NR)											40.21	24.81
GTP(NR)											29.33	12.79
PFK(NR)	13.38	22.71	9.58	23.81	10.03	40.76	11.47	13.28	19.65	13.87	29.31	41.19
TPI(NR)											44.43	74.19
GCK(NR)											43.88	47.24
GAPDH(NR)											50.97	57.67
SDHB(NR)											27.74	9.27
HKII(NR)	13.08	29.26	17.02	7.21	11.36	31.53					22.07	12.09

Table C.4: BCL-3(O) Gene (8 Introns) Versus Contrast Genes (Gray Cells Reject Ho at $\alpha = 0.01$ Level)

Contrast Gene	# Introns in Contrast Gene	# Introns Compared	Intron 1	Intron 2	Intron 3	Intron 4	Intron 5	Intron 6	Intron 7	Intron 8	Mean	SD
ABL(O)	10	8	214.60	38.50	261.01	31.47	37.43	146.11	60.11	251.40	130.08	100.61
BCR(O)	22	8	95.25	55.79	67.39	17.32	56.25	114.38	21.00	207.25	79.33	61.32
FMS(O)	21	8	97.08	40.18	70.34	36.17	61.34	264.35	83.89	81.25	91.83	72.82
FOS(O)	3	3	82.15	14.87	53.45						50.16	33.76
HST-1(O)	2	2	156.09	53.79							104.94	72.34
INT-2(O)	2	2	173.62	112.46							143.04	43.25
LCK(O)	11	8	13.64	29.55	42.21	11.35	19.25	30.43	12.70	83.59	30.34	24.06
MYC(O)	2	2	138.11	32.72							85.42	74.52
L-MYC(O)	2	2	151.02	39.61							95.32	78.78
N-MYC(O)	2	2	69.51	35.98							52.75	23.71
MAX(O)	1	1	40.15								40.15	NA
PIM1(O)	5	5	61.64	90.56	38.52	22.16	82.89				59.15	28.94
KIT(O)	20	8	44.03	24.97	76.89	53.44	92.63	75.34	32.66	195.50	74.43	54.18
RAFA1(O)	15	8	77.6	17.13	201.18	22.05	14.03	236.43	14.59	36.35	77.42	90.20
FPS(O)	18	8	73.1	52.18	118.84	12.66	46.21	10.11	19.38	41.92	46.80	36.25
WNT-1(O)	3	3	60.97	98.35	62.20						73.84	21.24
K-RAS2(O)	4	4	81.08	40.81	197.32	75.27					98.62	68.16
MLH1(TS)	18	8	83.25	22.73	127.95	40.76	70.25	129.72	47.64	212.87	91.90	62.41
MSH2(TS)	15	8	67.66	73.05	166.48	61.41	73.02	147.51	71.24	219.16	109.94	59.64
MTS1(TS)	3	3	34.18	67.49	106.29						69.32	36.09
RB(TS)	26	8	93.7	63.93	189.78	71.80	112.84	180.83	90.56	189.34	124.10	53.88
NF1(TS)	56	8	62.87	27.82	183.86	39.45	115.75	139.41	66.97	160.33	99.56	58.33
G6PD(NR)	12	8	58.22	36.28	49.76	21.18	24.22	34.84	23.87	75.45	40.48	19.22
PGK(NR)	10	8	69.28	27.46	87.91	44.61	52.15	80.67	45.97	111.31	64.92	27.48
ADOL(NR)	8	8	143.01	27.39	197.38	56.11	99.32	85.87	74.88	337.32	127.66	99.69
GTP(NR)	8	8	45.91	15.29	52.37	23.19	68.59	22.44	48.86	69.64	43.29	20.97
PFK(NR)	21	8	14.79	17.71	31.34	19.49	20.16	37.92	29.52	73.15	30.51	18.95
TPI(NR)	6	6	108.28	18.34	43.58	19.45	43.58	32.80			44.34	33.22
GCK(NR)	9	8	91.46	40.42	57.18	21.67	28.96	31.82	17.44	130.85	52.48	39.62
GAPDH(NR)	8	8	115.27	112.19	31.34	26.65	7.25	11.31	22.99	58.00	48.13	43.28
SDHB(NR)	7	7	50.05	26.78	112.75	38.01	40.13	69.92	30.14		52.54	30.18
HKII(NR)	17	8	40.52	18.41	63.85	32.55	29.93	53.27	18.48	85.14	42.77	23.32

Table C.5: FOS(O) Gene (3 Introns) Versus Contrast Genes (Gray Cells Reject Ho at α = 0.01 Level)

Contrast Gene	# Introns in Contrast Gene	# Introns Compared	Intron 1	Intron 2	Intron 3	Mean	SD
ABL(O)	10	3	222.63	30.15	13.44	88.74	116.25
BCR(O)	22	3	133.19	11.38	12.60	52.39	69.98
FMS(O)	21	3	122.27	4.51	22.29	49.69	63.48
BCL-3(O)	8	3	82.15	14.87	53.45	50.16	33.76
HST-1(O)	2	2	37.41	22.18		29.80	10.77
INT-2(O)	2	2	23.01	37.31		30.16	10.11
LCK(O)	11	3	39.5	20.36	33.78	31.21	9.82
MYC(O)	2	2	13.67	15.77		14.72	1.48
L-MYC(O)	2	2	37.87	9.92		23.90	19.76
N-MYC(O)	2	2	17.37	11.21		14.29	4.36
MAX(O)	1	1	41.94			41.94	NA
PIM1(O)	5	3	15.25	49.74	32.62	32.54	17.25
KIT(O)	20	3	11.93	20.25	14.71	15.63	4.24
RAFA1(O)	15	3	36.61	11.10	15.72	21.14	13.59
FPS(O)	18	3	34.39	29.96	19.48	27.94	7.66
WNT-1(O)	3	3	19.01	39.34	31.56	29.97	10.26
K-RAS2(O)	4	3	34.39	45.04	8.44	29.29	18.83
MLH1(TS)	18	3	29.25	22.71	9.04	20.33	10.31
MSH2(TS)	15	3	13.01	74.63	17.06	34.90	34.47
MTS1(TS)	3	3	41.86	25.00	14.20	27.02	13.94
RB(TS)	26	3	41.63	60.10	13.77	38.50	23.32
NF1(TS)	56	3	43.44	25.79	14.98	28.07	14.37
G6PD(NR)	12	3	14.67	10.10	30.65	18.47	10.79
PGK(NR)	10	3	31.66	23.30	19.03	24.66	6.42
ADOL(NR)	8	3	214.12	17.03	12.40	81.18	115.15
GTP(NR)	8	3	32.95	19.22	15.77	22.65	9.09
PFK(NR)	21	3	12.96	12.13	25.47	16.85	7.47
TPI(NR)	6	3	14.49	9.13	13.39	12.34	2.83
GCK(NR)	9	3	80.63	17.54	49.10	49.09	31.55
GAPDH(NR)	8	3	22.47	31.97	17.86	24.10	7.19
SDHB(NR)	7	3	36.74	20.88	19.80	25.81	9.48
HKII(NR)	17	3	22.06	10.00	17.59	16.55	6.10

Table C.6: HST-1(O) Gene (2 Introns) Versus Contrast Genes (Gray Cells Reject Ho at α = 0.01 Level)

Contrast Gene	# Introns in Contrast Gene	# Introns Compared	Intron 1	Intron 2	Mean	SD
ABL(O)	10	2	379.40	94.15	236.78	201.70
BCR(O)	22	2	269.46	19.55	144.50	176.72
FMS(O)	21	2	224.35	35.34	129.84	133.65
BCL-3(O)	8	2	156.09	53.79	104.94	72.34
FOS(O)	3	2	37.41	22.18	29.80	10.77
INT-2(O)	2	2	25.22	43.27	34.25	12.76
LCK(O)	11	2	67.6	15.05	41.33	37.16
MYC(O)	2	2	50.64	65.88	58.26	10.78
L-MYC(O)	2	2	10.45	57.51	33.98	33.28
N-MYC(O)	2	2	52.78	59.55	56.17	4.79
MAX(O)	1	1	88.29		88.29	NA
PIM1(O)	5	2	8.91	44.00	26.46	24.81
KIT(O)	20	2	45.72	59.08	52.40	9.45
RAFA1(O)	15	2	95.47	18.72	57.10	54.27
FPS(O)	18	2	30.36	31.33	30.85	0.69
WNT-1(O)	3	2	40.38	36.54	38.46	2.72
K-RAS2(O)	4	2	52.33	106.67	79.50	38.42
MLH1(TS)	18	2	64.71	71.53	68.12	4.82
MSH2(TS)	15	2	45.96	132.73	89.35	61.36
MTS1(TS)	3	2	81.37	13.21	47.29	48.20
RB(TS)	26	2	97.93	137.39	117.66	27.90
NF1(TS)	56	2	67.5	60.70	64.10	4.81
G6PD(NR)	12	2	56.58	42.73	49.66	9.79
PGK(NR)	10	2	81.82	71.31	76.57	7.43
ADOL(NR)	8	2	357.69	53.24	205.47	215.28
GTP(NR)	8	2	84.29	58.17	71.23	18.47
PFK(NR)	21	2	21.91	11.30	16.61	7.50
TPI(NR)	6	2	31.47	10.64	21.06	14.73
GCK(NR)	9	2	119.1	9.16	64.13	77.74
GAPDH(NR)	8	2	16.23	30.52	23.38	10.10
SDHB(NR)	7	2	108.4	63.51	85.96	31.74
HKII(NR)	17	2	38.73	29.52	34.13	6.51

Table C.7: INT-2(O) Gene (2 Introns) Versus Contrast Genes (Gray Cells Reject Ho at α = 0.01 Level)

Contrast Gene	# Introns in Contrast Gene	# Introns Compared	Intron 1	Intron 2	Mean	SD
ABL(O)	10	2	847.61	184.91	516.26	468.60
BCR(O)	22	2	512.28	131.21	321.75	269.45
FMS(O)	21	2	311.19	217.32	264.25	66.37
BCL-3(O)	8	2	173.62	112.46	143.04	43.25
FOS(O)	3	2	23.01	37.31	30.16	10.11
HST-1(O)	2	2	25.22	43.27	34.25	12.76
LCK(O)	11	2	71.50	13.52	42.51	41.00
MYC(O)	2	2	59.07	171.36	115.21	79.40
L-MYC(O)	2	2	22.37	240.66	131.52	154.35
N-MYC(O)	2	2	30.47	232.63	131.55	142.95
MAX(O)	1	1	58.84	███	58.84	NA
PIM1(O)	5	2	9.19	120.87	65.03	78.97
KIT(O)	20	2	30.11	59.77	44.94	20.97
RAFA1(O)	15	2	89.45	24.16	56.81	46.17
FPS(O)	18	2	24.34	34.35	29.35	7.08
WNT-1(O)	3	2	29.98	137.80	83.89	76.24
K-RAS2(O)	4	2	57.02	148.22	102.62	64.49
MLH1(TS)	18	2	65.16	87.18	76.17	15.57
MSH2(TS)	15	2	27.69	191.89	109.79	116.10
MTS1(TS)	3	2	68.03	66.12	67.07	1.35
RB(TS)	26	2	109.26	240.71	174.99	92.95
NF1(TS)	56	2	64.05	74.01	69.03	7.04
G6PD(NR)	12	2	32.55	243.56	138.05	149.20
PGK(NR)	10	2	69.54	94.78	82.16	17.84
ADOL(NR)	8	2	533.91	111.08	322.50	298.99
GTP(NR)	8	2	60.46	77.49	68.98	12.04
PFK(NR)	21	2	13.95	11.99	12.97	1.39
TPI(NR)	6	2	18.46	17.20	17.83	0.89
GCK(NR)	9	2	104.68	17.15	60.91	61.89
GAPDH(NR)	8	2	25.68	153.06	89.37	90.07
SDHB(NR)	7	2	101.75	67.92	84.84	23.92
HKII(NR)	17	2	26.24	48.63	37.44	15.83

Table C.8: LCK(O) Gene (11 Introns) Versus Contrast Genes (Gray Cells Reject Ho at α = 0.01 Level)

Contrast Gene	# Introns in Contrast Gene	# Introns Compared	Intron 1	Intron 2	Intron 3	Intron 4	Intron 5	Intron 6	Intron 7	Intron 8	Intron 9	Intron 10	Intron 11	Mean	SD
ABL(O)	10	10	61.64	47.79	62.24	47.63	51.60	47.05	32.09	52.07	12.35	8.98		42.34	18.70
BCR(O)	22	11	39.57	15.04	29.69	25.77	105.64	29.85	21.31	23.35	14.89	4.92	52.97	33.00	27.32
FMS(O)	21	11	39.62	26.81	30.17	32.63	93.63	67.97	40.60	23.30	7.78	7.15	26.67	36.03	25.27
BCL-3(O)	8	8	13.64	29.55	42.21	11.35	19.25	30.43	12.70	83.59				30.34	24.06
FOS(O)	3	3	39.50	20.36	33.78									31.21	9.82
HST-1(O)	2	2	67.60	15.05										41.33	37.16
INT-2(O)	2	2	71.50	13.52										42.51	41.00
MYC(O)	2	2	55.91	25.29										40.60	21.65
L-MYC(O)	2	2	80.27	25.03										52.65	39.06
N-MYC(O)	2	2	38.84	29.54										34.19	6.58
MAX(O)	1	1	31.44											31.44	NA
PIM1(O)	5	5	43.49	33.91	13.74	24.65	107.08							44.57	36.64
KIT(O)	20	11	32.34	34.80	49.62	60.02	101.41	30.05	29.65	77.06	39.98	18.69	10.86	44.04	26.56
RAFA1(O)	15	11	38.12	25.97	71.07	15.99	27.73	94.11	9.33	20.78	12.31	12.37	26.27	32.19	26.79
FPS(O)	18	11	47.21	20.12	33.52	13.62	89.13	15.62	19.50	25.84	25.96	5.27	38.93	30.43	22.83
WNT-1(O)	3	3	31.37	13.10	5.39									16.62	13.34
K-RAS2(O)	4	4	62.72	49.01	100.01	104.31								79.01	27.36
MLH1(TS)	18	11	56.24	43.91	68.31	46.55	83.66	57.06	45.49	68.92	25.89	13.89	31.06	49.18	20.54
MSH2(TS)	15	11	46.08	73.61	86.83	71.10	84.16	75.41	60.89	91.18	37.55	29.26	40.35	63.31	21.75
MTS1(TS)	3	3	32.25	18.13	54.42									34.93	18.29
RB(TS)	26	11	58.14	61.06	92.63	85.70	122.32	93.71	62.55	78.48	41.08	48.51	29.60	70.34	27.08
NF1(TS)	56	11	49.78	37.08	97.08	47.14	122.52	65.14	53.50	73.12	34.70	34.39	46.89	60.12	27.91
G6PD(NR)	12	11	33.06	18.75	23.82	29.07	42.27	14.08	14.19	33.68	8.06	8.13	31.03	23.29	11.43
PGK(NR)	10	10	46.68	36.46	41.99	45.32	71.97	37.14	46.28	38.98	34.41	14.31		41.35	14.27
ADOL(NR)	8	8	49.97	34.55	92.18	73.06	126.52	37.47	53.09	78.13				68.12	31.02
GTP(NR)	8	8	30.96	29.14	24.21	24.51	81.30	15.06	35.47	29.51				33.77	20.13
PFK(NR)	21	11	19.97	9.40	15.60	17.42	32.33	26.30	21.41	27.07	9.14	5.31	32.18	19.65	9.29
TPI(NR)	6	6	43.16	9.01	29.44	21.06	60.87	10.42						28.99	20.12
GCK(NR)	9	9	43.90	6.74	17.89	28.79	41.65	15.13	14.70	49.42	19.14			26.37	15.21
GAPDH(NR)	8	8	62.35	13.32	16.38	21.00	25.48	14.39	13.72	10.27				22.11	16.95
SDHB(NR)	7	7	31.99	34.71	51.40	36.92	59.21	32.77	24.70					38.81	12.09
HKII(NR)	17	11	38.53	16.08	25.37	22.39	44.87	29.31	20.47	21.82	7.12	11.26	31.41	24.42	11.22

Table C.9: MYC(O) Gene (2 Introns) Versus Contrast Genes (Gray Cells Reject Ho at α = 0.01 Level)

Contrast Gene	# Introns in Contrast Gene	# Introns Compared	Intron 1	Intron 2	Mean	SD
ABL(O)	10	2	431.10	31.57	231.33	282.51
BCR(O)	22	2	288.83	88.54	188.69	141.63
FMS(O)	21	2	165.74	47.03	106.39	83.94
BCL-3(O)	8	2	138.11	32.72	85.42	74.52
FOS(O)	3	2	13.67	15.77	14.72	1.48
HST-1(O)	2	2	50.64	65.88	58.26	10.78
INT-2(O)	2	2	59.07	171.36	115.21	79.40
LCK(O)	11	2	55.91	25.29	40.60	21.65
L-MYC(O)	2	2	42.96	21.19	32.08	15.39
N-MYC(O)	2	2	20.26	14.56	17.41	4.03
MAX(O)	1	1	69.00		69.00	NA
PIM1(O)	5	2	14.49	77.10	45.80	44.27
KIT(O)	20	2	13.45	19.49	16.47	4.27
RAFA1(O)	15	2	54.86	27.92	41.39	19.05
FPS(O)	18	2	53.23	42.24	47.74	7.77
WNT-1(O)	3	2	40.93	79.65	60.29	27.38
K-RAS2(O)	4	2	24.62	38.84	31.73	10.06
MLH1(TS)	18	2	24.76	19.47	22.12	3.74
MSH2(TS)	15	2	9.20	56.08	32.64	33.15
MTS1(TS)	3	2	65.72	69.34	67.53	2.56
RB(TS)	26	2	34.71	51.96	43.34	12.20
NF1(TS)	56	2	34.39	22.87	28.63	8.15
G6PD(NR)	12	2	23.41	45.22	34.32	15.42
PGK(NR)	10	2	26.54	20.36	23.45	4.37
ADOL(NR)	8	2	311.70	37.21	174.46	194.09
GTP(NR)	8	2	43.98	22.11	33.05	15.46
PFK(NR)	21	2	18.41	16.57	17.49	1.30
TPI(NR)	6	2	30.25	15.56	22.91	10.39
GCK(NR)	9	2	138.46	41.74	90.10	68.39
GAPDH(NR)	8	2	30.44	85.04	57.74	38.61
SDHB(NR)	7	2	42.25	19.21	30.73	16.29
HKII(NR)	17	2	27.86	8.43	18.15	13.74

Table C.10: L-MYC(O) Gene (2 Introns) Versus Contrast Genes (Gray Cells Reject Ho at α = 0.01 Level)

Contrast Gene	# Introns in Contrast Gene	# Introns Compared	Intron 1	Intron 2	Mean	SD
ABL(O)	10	2	269.55	31.83	150.69	168.10
BCR(O)	22	2	198.14	106.87	152.51	64.54
FMS(O)	21	2	184.71	39.05	111.88	102.99
BCL-3(O)	8	2	151.02	39.61	95.32	78.78
FOS(O)	3	2	37.87	9.92	23.90	19.76
HST-1(O)	2	2	10.45	57.51	33.98	33.28
INT-2(O)	2	2	22.37	240.66	131.52	154.35
LCK(O)	11	2	80.27	25.03	52.65	39.06
MYC(O)	2	2	42.96	21.19	32.08	15.39
N-MYC(O)	2	2	49.62	13.33	31.48	25.66
MAX(O)	1	1	76.15	■	76.15	NA
PIM1(O)	5	2	10.37	90.71	50.54	56.81
KIT(O)	20	2	40.94	22.57	31.76	12.99
RAFA1(O)	15	2	78.91	19.76	49.34	41.83
FPS(O)	18	2	32.08	37.65	34.87	3.94
WNT-1(O)	3	2	46.38	97.25	71.82	35.97
K-RAS2(O)	4	2	42.55	73.61	58.08	21.96
MLH1(TS)	18	2	53.82	23.65	38.74	21.33
MSH2(TS)	15	2	39.76	81.53	60.65	29.54
MTS1(TS)	3	2	73.48	58.28	65.88	10.75
RB(TS)	26	2	78.95	74.69	76.82	3.01
NF1(TS)	56	2	52.80	31.73	42.27	14.90
G6PD(NR)	12	2	49.69	62.01	55.85	8.71
PGK(NR)	10	2	65.06	26.11	45.59	27.54
ADOL(NR)	8	2	288.76	29.58	159.17	183.27
GTP(NR)	8	2	71.00	20.85	45.93	35.46
PFK(NR)	21	2	18.65	14.17	16.41	3.17
TPI(NR)	6	2	36.58	12.64	24.61	16.93
GCK(NR)	9	2	105.74	34.24	69.99	50.56
GAPDH(NR)	8	2	16.32	97.77	57.05	57.59
SDHB(NR)	7	2	-93.12	22.53	57.83	49.91
HKII(NR)	17	2	35.14	6.43	20.79	20.30

Table C.11: N-MYC(O) Gene (2 Introns) Versus Contrast Genes (Gray Cells Reject Ho at α = 0.01 Level)

Contrast Gene	# Introns in Contrast Gene	# Introns Compared	Intron 1	Intron 2	Mean	SD
ABL(O)	10	2	172.87	31.47	102.17	99.98
BCR(O)	22	2	75.70	100.87	88.29	17.80
FMS(O)	21	2	120.67	43.29	81.98	54.72
BCL-3(O)	8	2	69.51	35.98	52.75	23.71
FOS(O)	3	2	17.37	11.21	14.29	4.36
HST-1(O)	2	2	52.78	59.55	56.17	4.79
INT-2(O)	2	2	30.47	232.63	131.55	142.95
LCK(O)	11	2	38.84	29.54	34.19	6.58
MYC(O)	2	2	20.26	14.56	17.41	4.03
L-MYC(O)	2	2	49.62	13.33	31.48	25.66
MAX(O)	1	1	32.70		32.70	NA
PIM1(O)	5	2	16.60	72.53	44.56	39.55
KIT(O)	20	2	17.31	20.38	18.84	2.17
RAFA1(O)	15	2	26.13	19.86	23.00	4.43
FPS(O)	18	2	23.25	40.39	31.82	12.12
WNT-1(O)	3	2	24.19	83.57	53.88	41.99
K-RAS2(O)	4	2	36.26	54.69	45.48	13.03
MLH1(TS)	18	2	37.88	19.82	28.85	12.77
MSH2(TS)	15	2	11.16	71.24	41.20	42.48
MTS1(TS)	3	2	45.34	61.30	53.32	11.29
RB(TS)	26	2	43.75	63.88	53.82	14.23
NF1(TS)	56	2	37.25	27.16	32.21	7.14
G6PD(NR)	12	2	14.16	46.75	30.45	23.04
PGK(NR)	10	2	26.32	20.50	23.41	4.12
ADOL(NR)	8	2	182.91	31.78	107.34	106.86
GTP(NR)	8	2	25.32	13.80	19.56	8.15
PFK(NR)	21	2	12.20	13.26	12.73	0.75
TPI(NR)	6	2	13.66	14.81	14.24	0.81
GCK(NR)	9	2	50.45	39.59	45.02	7.68
GAPDH(NR)	8	2	38.90	88.59	63.74	35.14
SDHB(NR)	7	2	28.91	20.72	24.82	5.79
HKII(NR)	17	2	20.88	7.47	14.18	9.48

Table C.12: MAX(O) Gene (1 Intron) Versus Contrast Genes (Gray Cells Reject Ho at $\alpha = 0.01$ Level)

Contrast Gene	# Introns in Contrast Gene	# Introns Compared	Intron 1
ABL(O)	10	1	94.49
BCR(O)	22	1	31.10
FMS(O)	21	1	71.42
BCL-3(O)	8	1	40.15
FOS(O)	3	1	41.94
HST-1(O)	2	1	88.29
INT-2(O)	2	1	58.84
LCK(O)	11	1	31.44
MYC(O)	2	1	69.00
N-MYC(O)	2	1	76.15
L-MYC	2	1	32.70
PIM1(O)	5	1	41.99
KIT(O)	20	1	30.03
RAFA1(O)	15	1	39.54
FPS(O)	18	1	32.11
WNT-1(O)	3	1	35.20
K-RAS2(O)	4	1	68.23
MLH1(TS)	18	1	55.82
MSH2(TS)	15	1	36.28
MTS1(TS)	3	1	30.41
RB(TS)	26	1	67.68
NF1(TS)	56	1	49.91
G6PD(NR)	12	1	27.73
PGK(NR)	10	1	39.30
ADOL(NR)	8	1	72.53
GTP(NR)	8	1	15.16
PFK(NR)	21	1	10.88
TPI(NR)	6	1	52.10
GCK(NR)	9	1	18.70
GAPDH(NR)	8	1	66.40
SDHB(NR)	7	1	37.39
HKII(NR)	17	1	24.49

Table C.13: PIM1(O) Gene (5 Introns) Versus Contrast Genes (Gray Cells Reject Ho at α = 0.01 Level)

Contrast Gene	# Introns in Contrast Gene	# Introns Compared	Intron 1	Intron 2	Intron 3	Intron 4	Intron 5	Mean	SD
ABL(O)	10	5	84.47	66.66	46.34	49.14	29.33	55.19	21.05
BCR(O)	22	5	61.08	85.61	26.68	68.32	54.33	59.20	21.60
FMS(O)	21	5	83.06	96.68	27.49	80.43	29.40	63.41	32.52
BCL-3	8	5	61.64	90.56	38.52	22.16	82.89	59.15	28.94
FOS(O)	3	3	15.25	49.74	32.62			32.54	17.25
HST-1(O)	2	2	8.91	44.00				26.46	24.81
INT-2(O)	2	2	9.19	120.87				65.03	78.97
LCK(O)	11	5	43.49	33.91	13.74	24.65	107.08	44.57	36.64
MYC(O)	2	2	14.49	77.10				45.80	44.27
L-MYC(O)	2	2	10.37	90.71				50.54	56.81
N-MYC(O)	2	2	16.60	72.53				44.56	39.55
MAX(O)	1	1	41.99					41.99	NA
KIT(O)	20	5	22.66	56.04	58.45	53.91	26.99	43.61	17.29
RAFA1(O)	15	5	29.48	28.10	52.96	27.06	23.79	32.28	11.75
FPS(O)	18	5	12.96	41.12	30.34	16.98	44.49	29.18	14.06
WNT-1(O)	3	3	18.09	24.78	18.90			20.59	3.65
K-RAS2(O)	4	4	15.69	85.22	70.45	129.66		75.26	47.01
MLH1(TS)	18	5	29.28	58.38	59.41	48.09	19.13	42.86	17.96
MSH2(TS)	15	5	18.10	78.28	74.98	88.41	37.01	59.36	30.20
MTS1(TS)	3	3	43.42	50.43	47.65			47.17	3.53
RB(TS)	26	5	32.98	81.14	68.76	100.67	39.27	64.57	28.43
NF1(TS)	56	5	39.85	56.26	68.99	23.08	46.26	46.89	17.27
G6PD(NR)	12	5	22.95	74.84	13.72	74.90	86.23	54.52	33.52
PGK(NR)	10	5	30.47	62.50	51.48	36.83	19.94	40.24	16.89
ADOL(NR)	8	5	122.17	101.20	75.60	67.66	16.94	76.71	39.76
GTP(NR)	8	5	36.99	59.17	22.40	33.79	18.91	34.25	15.85
PFK(NR)	21	5	13.62	24.17	19.57	51.36	35.75	28.90	14.95
TPI(NR)	6	5	9.98	24.01	35.10	25.34	23.69	23.63	8.96
GCK(NR)	9	5	48.28	43.07	15.55	49.30	45.30	40.30	14.05
GAPDH(NR)	8	5	8.82	32.93	10.33	25.55	24.48	20.42	10.44
SDHB(NR)	7	5	38.53	46.52	46.43	19.13	12.67	32.65	15.80
HKII(NR)	17	5	26.66	37.24	29.67	25.93	33.11	30.52	4.71

Table C.14: KIT(O) Gene (20 Introns) Versus Contrast Genes (Gray Cells Reject Ho at α= 0.01 Level)

Contrast Gene	# Introns in Contrast Gene	# Introns Compared	Intron 1	Intron 2	Intron 3	Intron 4	Intron 5	Intron 6	Intron 7	Intron 8	Intron 9	Intron 10	Intron 11
ABL(O)	10	10	43.83	11.76	22.73	52.26	52.66	25.07	11.25	34.46	34.69	17.93	
BCR(O)	22	20	31.78	44.59	26.77	108.42	100.57	44.54	19.61	187.17	98.28	26.80	67.27
FMS(O)	20	20	42.14	32.43	35.02	93.57	69.91	72.35	13.05	62.57	46.69	30.67	48.66
BCL-3(O)	8	8	44.03	24.97	76.89	53.44	92.63	75.34	32.66	195.50			
FOS(O)	3	3	11.93	20.25	14.71								
HST-1(O)	2	2	45.72	59.08									
INT-2(O)	2	2	30.11	59.77									
LCK(O)	11	11	32.34	34.80	49.62	60.02	101.41	30.05	29.65	77.06	39.98	18.69	10.86
MYC(O)	2	2	13.45	19.49									
L-MYC(O)	2	2	40.94	22.57									
N-MYC(O)	2	2	17.31	20.38									
MAX(O)	1	1	30.03										
PIM1(O)	5	5	22.66	56.04	58.45	53.91	26.99						
RAFA1(O)	15	15	13.18	23.40	20.91	42.17	49.17	69.20	14.57	60.53	36.62	12.76	29.74
FPS(O)	18	18	38.07	40.09	33.09	45.89	71.64	21.59	29.56	72.61	82.98	30.45	58.24
WNT-1(O)	3	3	14.80	60.10	48.77								
K-RAS2(O)	4	4	19.22	13.26	22.08	19.96							
MLH1(TS)	18	18	7.67	11.72	14.87	9.11	9.93	35.21	16.09	30.27	22.66	10.54	36.69
MSH2(TS)	15	15	9.41	33.80	15.91	25.39	17.75	55.72	19.88	16.36	10.23	9.87	23.95
MTS1(TS)	3	3	32.49	51.28	23.53								
RB(TS)	26	20	13.38	14.61	17.70	13.69	12.40	74.87	23.66	13.37	21.06	18.57	20.72
NF1(TS)	56	20	22.27	14.16	22.82	18.44	6.67	41.84	14.02	14.44	17.04	11.49	33.49
G6PD(NR)	12	12	6.20	34.02	53.21	100.63	118.58	40.65	21.45	126.99	62.54	35.29	36.43
PGK(NR)	10	10	11.29	8.08	26.59	15.08	16.77	25.53	17.88	26.36	25.78	19.66	
ADOL(NR)	8	8	72.98	18.94	12.55	29.47	20.70	20.60	17.41	44.66			
GTP(NR)	8	8	18.19	16.03	25.03	43.50	14.37	17.84	7.92	43.75			
PFK(NR)	21	20	17.62	20.33	36.64	103.35	63.19	28.28	75.10	67.95	42.78	43.50	41.54
TPI(NR)	6	6	27.03	29.80	15.44	50.88	52.49	22.99					
GCK(NR)	9	9	44.07	45.90	69.71	72.68	72.96	28.14	26.78	115.86	89.43		
GAPDH(NR)	8	8	31.58	52.65	33.95	37.52	43.49	15.52	22.31	32.57			
SDHB(NR)	7	7	18.95	10.94	27.39	16.95	37.14	21.67	8.22				
HKII(NR)	17	17	28.00	16.79	25.83	37.24	52.88	29.46	14.44	58.78	46.83	16.46	47.10

Table C.14 (Continued): KIT(O) Gene (20 Introns) Versus Contrast Genes (Gray Cells Reject Ho at α= 0.01 Level)

Contrast Gene	Intron 12	Intron 13	Intron 14	Intron 15	Intron 16	Intron 17	Intron 18	Intron 19	Intron 20	Mean	SD
ABL(O)										30.66	15.44
BCR(O)	32.39	70.54	64.80	32.94	125.08	27.12	22.99	67.66	35.17	61.72	43.36
FMS(O)	25.66	84.82	56.60	37.76	56.49	29.92	23.62	58.48	7.08	46.38	23.04
BCL-3(O)										74.43	54.18
FOS(O)										15.63	4.24
HST-1(O)										52.40	9.45
INT-2(O)										44.94	20.97
LCK(O)										44.04	26.56
MYC(O)										16.47	4.27
L-MYC(O)										31.76	12.99
N-MYC(O)										18.84	2.17
MAX(O)										30.03	NA
PIM1(O)										43.61	17.29
RAFA1(O)	30.42	76.98	51.17	60.29						39.41	21.08
FPS(O)	29.63	85.44	55.50	50.91	84.73	50.91	19.49			50.05	21.86
WNT-1(O)										41.22	23.57
K-RAS2(O)										18.63	3.78
MLH1(TS)	12.38	17.31	15.92	15.61	16.94	12.26	17.95			17.40	8.56
MSH2(TS)	15.01	7.72	11.17	18.44						19.37	12.27
MTS1(TS)										35.77	14.17
RB(TS)	19.04	13.78	26.18	47.42	15.22	11.54	23.73	16.53	28.16	22.28	14.80
NF1(TS)	17.52	11.00	11.55	27.61	11.84	6.62	14.61	10.42	18.91	17.34	8.82
G6PD(NR)	39.62									56.30	38.69
PGK(NR)										19.30	6.67
ADOL(NR)										29.66	20.08
GTP(NR)										23.33	13.38
PFK(NR)	29.12	47.56	28.57	33.17	31.34	45.78	23.13	32.77	18.77	41.52	21.55
TPI(NR)										33.11	15.19
GCK(NR)										62.84	29.39
GAPDH(NR)										33.70	11.56
SDHB(NR)										20.18	9.85
HKII(NR)	12.50	26.98	30.61	16.37	37.17	8.78				29.78	15.01

Table C.15: RAFA1(O) Gene (15 Introns) Versus Contrast Genes (Gray Cells Reject Ho at α= 0.01 Level)

Contrast Gene	# Introns in Contrast Gene	# Introns Compared	Intron 1	Intron 2	Intron 3	Intron 4	Intron 5	Intron 6	Intron 7	Intron 8	Intron 9	Intron 10	Intron 11
ABL(O)	10	10	160.74	24.67	25.69	25.42	20.91	109.17	9.97	38.01	14.78	45.79	
BCR(O)	22	15	63.84	15.18	20.88	18.36	14.84	201.33	9.07	22.44	6.78	66.98	19.65
FMS(O)	20	15	89.20	13.50	36.42	14.77	10.57	219.70	14.09	25.60	9.16	89.00	18.83
BCL-3(O)	8	8	77.60	17.13	201.18	22.05	14.03	236.43	14.59	36.35			
FOS(O)	3	3	36.61	11.10	15.72								
HST-1(O)	2	2	95.47	18.72									
INT-2(O)	2	2	89.45	24.16									
LCK(O)	11	11	38.12	25.97	71.07	15.99	27.73	94.11	9.33	20.78	12.31	12.37	26.27
MYC(O)	2	2	54.86	27.92									
L-MYC(O)	2	2	78.91	19.76									
N-MYC(O)	2	2	26.13	19.86									
MAX(O)	1	1	39.54										
PIM1(O)	5	5	29.48	28.1	52.96	27.06	23.79						
KIT(O)	20	15	13.18	23.40	20.91	42.17	49.17	69.20	14.57	60.53	36.62	12.76	29.74
FPS(O)	18	15	61.25	31.52	48.01	12.23	11.33	51.76	8.74	9.46	18.62	59.42	12.28
WNT-1(O)	3	3	27.49	32.93	97.95								
K-RAS2(O)	4	4	49.21	40.44	62.29	55.10							
MLH1(TS)	18	15	34.42	19.89	23.11	34.47	37.23	26.21	25.89	71.18	25.42	19.73	42.56
MSH2(TS)	15	15	20.37	69.19	36.24	51.01	47.33	33.21	34.60	68.25	29.27	54.91	48.19
MTS1(TS)	3	3	44.03	16.38	24.33								
RB(TS)	26	15	42.50	42.34	62.20	59.05	54.83	35.05	38.69	54.00	47.89	70.02	38.37
NF1(TS)	56	15	39.90	32.82	55.10	29.57	57.75	23.21	29.67	64.56	39.06	35.47	57.25
G6PD(NR)	12	12	7.40	18.17	38.18	11.98	14.36	92.29	4.54	17.90	9.86	29.82	13.18
PGK(NR)	10	10	18.20	28.64	33.34	24.59	25.23	18.09	24.51	42.59	27.64	19.64	
ADOL(NR)	8	8	159.82	26.11	44.35	31.98	33.05	28.65	23.40	57.20			
GTP(NR)	8	8	18.18	16.36	23.21	18.48	35.18	45.94	14.13	37.21			
PFK(NR)	21	15	17.35	11.78	30.54	16.48	10.61	85.44	9.36	18.90	5.03	65.21	12.22
TPI(NR)	6	6	58.84	15.06	16.26	16.11	12.13	56.29					
GCK(NR)	9	9	65.71	19.86	85.90	16.98	14.15	85.78	7.37	27.68	13.91		
GAPDH(NR)	8	8	74.82	29.38	19.04	12.65	11.33	55.92	2.78	22.46			
SDHB(NR)	7	7	29.80	21.24	54.21	22.67	14.14	53.57	13.52				
HKII(NR)	17	15	24.67	17.70	15.51	9.84	17.55	85.57	8.19	22.64	4.42	13.18	18.91

Table C.15 (Continued): RAFA1(O) Gene (15 Introns) Versus Contrast Genes (Gray Cells Reject Ho at α= 0.01 Level)

Contrast Gene	Intron 12	Intron 13	Intron 14	Intron 15	Mean	SD
ABL(O)	43.31	34.56	67.63	85.29	47.51	48.75
BCR(O)					46.01	49.64
FMS(O)	11.18	24.53	46.46	83.04	47.07	55.95
BCL-3(O)					77.42	90.20
FOS(O)					21.14	13.59
HST-1(O)					57.10	54.27
INT-2(O)					56.81	46.17
LCK(O)					32.19	26.79
MYC(O)					41.39	19.05
L-MYC(O)					49.34	41.83
N-MYC(O)					23.00	4.43
MAX(O)					39.54	NA
PIM1(O)					32.28	11.75
KIT(O)	30.42	76.98	51.17	60.29	39.41	21.08
FPS(O)	23.79	34.86	44.21	46.02	31.57	19.10
WNT-1(O)					52.79	39.20
K-RAS2(O)					51.76	9.25
MLH1(TS)	52.51	47.20	48.70	60.43	37.93	15.56
MSH2(TS)	61.51	73.62	17.51	97.56	49.52	21.97
MTS1(TS)					28.25	14.23
RB(TS)	122.59	77.77	54.46	93.43	59.54	23.69
NF1(TS)	59.76	76.42	104.20	91.09	53.05	23.76
G6PD(NR)	14.25				22.66	23.84
PGK(NR)					26.25	7.50
ADOL(NR)					50.57	45.49
GTP(NR)					26.09	11.75
PFK(NR)	16.84	14.84	22.92	19.13	23.78	22.11
TPI(NR)					29.12	22.10
GCK(NR)					37.48	32.23
GAPDH(NR)					28.55	24.59
SDHB(NR)					29.88	17.30
HKII(NR)	17.00	26.74	33.76	33.64	23.29	19.20

Table C.16: FPS(O) Gene (18 Introns) Versus Contrast Genes (Gray Cells Reject Ho at α= 0.01 Level)

Contrast Gene	# Introns in Contrast Gene	# Introns Compared	Intron 1	Intron 2	Intron 3	Intron 4	Intron 5	Intron 6	Intron 7	Intron 8	Intron 9	Intron 10	Intron 11
ABL(O)	10	10	197.03	60.92	46.7	21.61	30.6	25.78	39.32	57.35	38.27	39.88	
BCR(O)	22	18	100.64	32.95	7.11	11.78	52.30	16.49	15.49	36.31	35.05	24.45	20.46
FMS(O)	21	18	156.56	37.16	9.50	15.23	17.78	34.13	56.03	28.61	18.38	17.97	29.03
BCL-3(O)	8	8	73.1	52.18	118.84	12.66	46.21	10.11	19.38	41.92			
FOS(O)	3	3	34.39	29.96	19.48								
HST-1(O)	2	2	30.36	31.33									
INT-2(O)	2	2	24.34	34.35									
LCK(O)	11	11	47.21	20.12	33.52	13.62	89.13	15.62	19.50	25.84	25.96	5.27	38.93
MYC(O)	2	2	53.23	42.24									
L-MYC(O)	2	2	32.08	37.65									
N-MYC(O)	2	2	23.25	40.39									
MAX(O)	1	1	32.11										
PIM1(O)	5	5	12.96	41.12	30.34	16.98	44.49						
KIT(O)	20	18	38.07	40.09	33.09	45.89	71.64	21.59	29.56	72.61	82.98	30.45	58.24
RAFA1(O)	15	15	61.25	31.52	48.01	12.23	11.33	51.76	8.74	9.46	18.62	59.42	12.28
WNT-1(O)	3	3	27.95	39.95	42.60								
K-RAS2(O)	4	4	48.48	64.23	108.96	66.30							
MLH1(TS)	18	18	65.82	54.70	46.88	35.11	52.41	46.47	47.76	97.91	71.33	48.78	71.19
MSH2(TS)	15	15	35.54	83.33	78.68	59.44	74.13	66.73	60.31	86.15	79.85	107.85	116.19
MTS1(TS)	3	3	50.92	25.06	19.16								
RB(TS)	26	18	85.54	76.72	97.04	69.37	93.64	66.80	82.99	72.39	84.34	150.42	98.54
NF1(TS)	56	18	63.64	46.39	96.28	23.87	113.64	59.55	57.55	70.14	78.63	62.59	137.69
G6PD(NR)	12	12	40.29	40.98	24.04	4.94	53.11	9.23	17.84	31.89	15.97	15.95	15.25
PGK(NR)	10	10	63.25	43.64	32.90	31.93	29.35	34.20	48.37	58.99	31.08	40.66	
ADOL(NR)	8	8	202.28	44.98	69.58	37.07	58.32	30.13	55.95	88.89			
GTP(NR)	8	8	46.91	42.99	14.40	18.24	45.11	8.92	42.94	49.48			
PFK(NR)	21	18	8.47	8.19	16.74	11.98	20.32	11.65	15.33	21.16	24.82	10.78	21.68
TPI(NR)	6	6	20.41	19.80	18.53	7.18	17.51	13.34					
GCK(NR)	9	9	44.05	24.44	41.53	11.21	35.44	6.15	24.71	30.68	27.73		
GAPDH(NR)	8	8	24.87	26.71	9.50	11.66	10.14	7.66	12.23	23.59			
SDHB(NR)	7	7	69.51	46.75	27.47	23.50	8.78	23.52	34.89				
HKII(NR)	17	17	26.40	25.30	9.13	14.63	27.19	20.20	10.49	36.79	27.87	24.49	12.72

Table C.16 (Continued): FPS(O) Gene (18 Introns) Versus Contrast Genes (Gray Cells Reject Ho at α= 0.01 Level)

Contrast Gene	Intron 12	Intron 13	Intron 14	Intron 15	Intron 16	Intron 17	Intron 18	Mean	SD
ABL(O)								55.75	51.20
BCR(O)	7.06	22.66	18.23	29.55	23.24	34.10	26.63	28.58	21.27
FMS(O)	5.95	40.93	23.97	34.01	7.88	37.00	14.97	32.50	33.64
BCL-3(O)								46.80	36.25
FOS(O)								27.94	7.66
HST-1(O)								30.85	0.69
INT-2(O)								29.35	7.08
LCK(O)								30.43	22.83
MYC(O)								47.74	7.77
L-MYC(O)								34.87	3.94
N-MYC(O)								31.82	12.12
MAX(O)								32.11	NA
PIM1(O)								29.18	14.06
KIT(O)	29.63	85.44	55.50	50.91	84.73	50.91	19.49	50.05	21.86
RAFA1(O)	23.79	34.86	44.21	46.02				31.57	19.10
WNT-1(O)								36.83	7.81
K-RAS2(O)								71.99	25.90
MLH1(TS)	35.97	47.61	31.98	33.98	26.47	55.61	16.42	49.25	19.12
MSH2(TS)	76.03	91.90	25.83	71.59				74.24	23.60
MTS1(TS)								31.71	16.89
RB(TS)	122.01	93.15	36.55	72.40	86.63	84.01	118.64	88.40	24.69
NF1(TS)	69.29	54.74	81.50	79.03	60.10	66.10	88.77	72.75	25.52
G6PD(NR)	13.33							23.57	14.77
PGK(NR)								41.44	12.05
ADOL(NR)								73.40	55.29
GTP(NR)								33.62	16.69
PFK(NR)	9.31	15.92	11.39	15.87	8.25	15.86	14.98	14.59	5.02
TPI(NR)								16.13	5.05
GCK(NR)								27.33	12.66
GAPDH(NR)								15.79	7.84
SDHB(NR)								33.49	19.67
HKII(NR)	11.71	35.97	9.69	31.22	12.63	35.25		21.86	9.87

Table C.17: WNT-1(O) Gene (3 Introns) Versus Contrast Genes (Gray Cells Reject Ho at α = 0.01 Level)

Contrast Gene	# Introns in Contrast Gene	# Introns Compared	Intron 1	Intron 2	Intron 3	Mean	SD
ABL(O)	10	3	157.15	94.77	91.46	114.46	37.01
BCR(O)	22	3	84.48	90.38	28.86	67.91	33.94
FMS(O)	21	3	108.20	116.51	33.71	86.14	45.60
BCL-3(O)	8	3	60.97	98.35	62.20	73.84	21.24
FOS(O)	3	3	19.01	39.34	31.56	29.97	10.26
HST-1(O)	2	2	40.38	36.54		38.46	2.72
INT-2(O)	2	2	29.98	137.80		83.89	76.24
LCK(O)	11	3	31.37	13.10	5.39	16.62	13.34
MYC(O)	2	2	40.93	79.65		60.29	27.38
L-MYC(O)	2	2	46.38	97.25		71.82	35.97
N-MYC(O)	2	2	24.19	83.57		53.88	41.99
MAX(O)	1	1	35.2			35.20	NA
PIM1(O)	5	3	18.09	24.78	18.90	20.59	3.65
KIT(O)	20	3	14.8	60.10	48.77	41.22	23.57
RAFA1(O)	15	3	27.49	32.93	97.95	52.79	39.20
FPS(O)	18	3	27.95	39.95	42.60	36.83	7.81
K-RAS2(O)	4	3	40.70	116.48	115.00	90.72	43.33
MLH1(TS)	18	3	35.44	70.57	72.37	59.46	20.82
MSH2(TS)	15	3	23.28	115.26	96.70	78.41	48.64
MTS1(TS)	3	3	57.77	36.46	64.80	53.01	14.76
RB(TS)	26	3	59.12	133.86	104.71	99.23	37.67
NF1(TS)	56	3	52.20	60.46	111.95	74.87	32.38
G6PD(NR)	12	3	13.89	76.72	33.98	41.53	32.09
PGK(NR)	10	3	37.11	68.43	48.38	51.31	15.86
ADOL(NR)	8	3	169.16	84.46	114.08	122.57	42.98
GTP(NR)	8	3	39.26	60.48	18.84	39.53	20.82
PFK(NR)	21	3	14.82	13.89	23.08	17.26	5.06
TPI(NR)	6	3	25.23	15.31	23.58	21.37	5.32
GCK(NR)	9	3	40.13	17.59	27.86	28.53	11.29
GAPDH(NR)	8	3	43.85	14.34	17.85	25.35	16.12
SDHB(NR)	7	3	55.65	59.39	55.16	56.73	2.31
HKII(NR)	17	3	24.34	22.69	30.38	25.81	4.05

Table C.18: K-RAS2(O) Gene (4 Introns) Versus Contrast Genes (Gray Cells Reject Ho at α = 0.01 Level)

Contrast Gene	# Introns in Contrast Gene	# Introns Compared	Intron 1	Intron 2	Intron 3	Intron 4	Mean	SD
ABL(O)	10	4	122.13	33.07	67.86	140.82	90.97	49.47
BCR(O)	22	4	101.01	117.92	41.40	296.82	139.29	110.03
FMS(O)	21	4	104.41	91.08	79.54	213.12	122.04	61.57
BCL-3	8	4	81.08	40.81	197.32	75.27	98.62	68.16
FOS(O)	3	3	34.39	45.04	8.44		29.29	18.83
HST-1(O)	2	2	52.33	106.67			79.50	38.42
INT-2(O)	2	2	57.02	148.22			102.62	64.49
LCK(O)	11	4	62.72	49.01	100.01	104.31	79.01	27.36
MYC(O)	2	2	24.62	38.84			31.73	10.06
L-MYC(O)	2	2	42.55	73.61			58.08	21.96
N-MYC(O)	2	2	36.26	54.69			45.48	13.03
MAX(O)	1	1	68.23				68.23	NA
PIM1	5	4	15.69	85.22	70.45	129.66	75.26	47.01
KIT(O)	20	4	19.22	13.26	22.08	19.96	18.63	3.78
RAFA1(O)	15	4	49.21	40.44	62.29	55.10	51.76	9.25
FPS(O)	18	4	48.48	64.23	108.96	66.30	71.99	25.90
WNT-1(O)	3	3	40.70	116.48	115.00		90.72	43.33
MLH1(TS)	18	4	18.04	15.94	12.08	26.58	18.16	6.13
MSH2(TS)	15	4	18.87	32.27	15.53	31.64	24.58	8.63
MTS1(TS)	3	3	58.07	107.75	64.93		76.92	26.92
RB(TS)	26	4	22.56	20.30	13.00	16.01	17.97	4.28
NF1(TS)	56	4	30.94	15.10	11.05	24.52	20.40	9.01
G6PD(NR)	12	4	37.90	84.42	66.83	178.38	91.88	60.77
PGK(NR)	10	4	35.64	13.70	46.70	20.43	29.12	14.89
ADOL(NR)	8	4	168.10	57.64	46.29	22.52	73.64	64.65
GTP(NR)	8	4	53.35	21.64	31.15	63.15	42.32	19.22
PFK(NR)	21	4	23.69	29.20	62.56	153.60	67.26	60.07
TPI(NR)	6	4	46.56	38.42	23.63	85.06	48.42	26.21
GCK(NR)	9	4	102.35	80.17	120.77	108.37	102.91	16.99
GAPDH(NR)	8	4	25.46	135.19	36.63	55.87	63.29	49.56
SDHB(NR)	7	4	49.35	21.60	67.86	44.14	45.74	19.04
HKII(NR)	17	4	39.45	29.74	57.08	66.10	48.09	16.50

Table C.19: MLH1(TS) Gene (18 Introns) Versus Contrast Genes (Gray Cells Reject Ho at α= 0.01 Level)

Contrast Gene	# Introns in Contrast Gene	# Introns Compared	Intron 1	Intron 2	Intron 3	Intron 4	Intron 5	Intron 6	Intron 7	Intron 8	Intron 9	Intron 10	Intron 11
ABL(O)	10	10	60.62	6.09	21.83	48.87	41.68	31.39	25.89	33.16	20.01	19.55	
BCR(O)	22	18	61.40	59.99	25.73	80.51	83.17	63.41	44.69	140.67	53.63	33.27	86.66
FMS(O)	21	18	57.32	35.04	43.05	73.40	53.51	57.67	24.42	49.73	32.10	36.06	89.11
BCL-3(O)	8	8	83.25	22.73	127.95	40.76	70.25	129.72	47.64	212.87			
FOS(O)	3	3	29.25	22.71	9.04								
HST-1(O)	2	2	64.71	71.53									
INT-2(O)	2	2	65.16	87.18									
LCK(O)	11	11	56.24	43.91	68.31	46.55	83.66	57.06	45.49	68.92	25.89	13.89	31.06
MYC(O)	2	2	24.76	19.47									
L-MYC(O)	2	2	53.82	23.65									
N-MYC(O)	2	2	37.88	19.82									
MAX(O)	1	1	55.82										
PIM1(O)	5	5	29.28	58.38	59.41	48.09	19.13						
KIT(O)	20	18	7.67	11.72	14.87	9.11	9.93	35.21	16.09	30.27	22.66	10.54	36.69
RAFA1(O)	15	15	34.42	19.89	23.11	34.47	37.23	26.21	25.89	71.18	25.42	19.73	42.56
FPS(O)	18	18	65.82	54.70	46.88	35.11	52.41	46.47	47.76	97.91	71.33	48.78	71.19
WNT-1(O)	3	3	35.44	70.57	72.37								
K-RAS2(O)	4	4	18.04	15.94	12.08	26.58							
MSH2(TS)	15	15	11.15	29.71	9.16	23.94	12.96	34.83	10.46	28.77	20.58	29.50	30.25
MTS1(TS)	3	3	50.84	66.85	27.06								
RB(TS)	26	18	7.43	11.05	10.96	17.76	8.85	28.38	13.46	40.08	31.89	49.42	20.22
NF1(TS)	56	18	20.35	13.33	12.46	22.82	9.36	11.86	13.36	30.22	25.03	33.88	34.63
G6PD(NR)	12	12	21.40	38.85	47.92	72.09	86.08	68.36	42.33	107.01	44.86	27.44	40.62
PGK(NR)	10	10	16.68	7.56	29.36	21.88	20.40	7.37	11.17	15.42	34.32	24.80	
ADOL(NR)	8	8	99.45	26.75	13.21	32.05	16.50	11.47	29.22	38.95			
GTP(NR)	8	8	35.39	11.34	22.39	44.99	10.45	30.83	14.39	37.47			
PFK(NR)	21	18	27.68	27.45	41.09	79.87	50.37	56.14	88.76	57.67	27.51	51.44	47.98
TPI(NR)	6	6	57.76	25.58	10.10	41.87	46.59	39.42					
GCK(NR)	9	9	80.27	54.30	85.85	59.18	57.90	57.52	36.84	107.25	59.70		
GAPDH(NR)	8	8	41.91	73.26	29.14	33.38	31.83	43.56	32.28	33.36			
SDHB(NR)	7	7	32.63	7.02	29.93	18.40	26.84	28.06	23.91				
HKII(NR)	17	17	31.67	15.73	26.62	30.15	52.77	57.90	30.95	42.36	24.15	21.59	48.78

Table C.19 (Continued): MLH1(TS) Gene (18 Introns) Versus Contrast Genes (Gray Cells Reject Ho at α= 0.01 Level)

Contrast Gene	Intron 12	Intron 13	Intron 14	Intron 15	Intron 16	Intron 17	Intron 18	Mean	SD
ABL(O)	46.90	38.50	30.02	17.93	11.42	34.11	20.23	30.91	15.98
BCR(O)	29.04	38.16	26.62	25.28	18.35	33.49	21.50	51.79	31.67
FMS(O)								41.33	18.96
BCL-3(O)								91.90	62.41
FOS(O)								20.33	10.31
HST-1(O)								68.12	4.82
INT-2(O)								76.17	15.57
LCK(O)								49.18	20.54
MYC(O)								22.12	3.74
L-MYC(O)								38.74	21.33
N-MYC(O)								28.85	12.77
MAX(O)								55.82	NA
PIM1(O)								42.86	17.96
KIT(O)	12.38	17.31	15.92	15.61	16.94	12.26	17.95	17.40	8.56
RAFA1(O)	52.51	47.20	48.70	60.43				37.93	15.56
FPS(O)	35.97	47.61	31.98	33.98	26.47	55.61	16.42	49.25	19.12
WNT-1(O)								59.46	20.82
K-RAS2(O)								18.16	6.13
MSH2(TS)	28.31	23.38	21.18	15.95				22.01	8.35
MTS1(TS)								48.25	20.02
RB(TS)	37.01	18.27	20.54	35.75	17.34	12.02	47.30	23.76	13.45
NF1(TS)	25.41	21.88	40.36	22.72	14.85	14.00	41.17	22.65	10.03
G6PD(NR)	37.25							52.85	25.35
PGK(NR)								18.90	9.01
ADOL(NR)								33.45	28.36
GTP(NR)								25.91	13.14
PFK(NR)	24.52	30.70	18.35	25.09	17.25	41.42	17.63	40.61	20.78
TPI(NR)								36.89	16.77
GCK(NR)								66.53	20.94
GAPDH(NR)								39.84	14.42
SDHB(NR)								23.83	8.70
HKII(NR)	7.00	20.25	18.54	10.65	8.29	20.25		27.51	15.24

Table C.20: MSH2(TS) Gene (15 Introns) Versus Contrast Genes (Gray Cells Reject Ho at α= 0.01 Level)

Contrast Gene	# Introns in Contrast Gene	# Introns Compared	Intron 1	Intron 2	Intron 3	Intron 4	Intron 5	Intron 6	Intron 7	Intron 8	Intron 9	Intron 10	Intron 11
ABL(O)	10	10	69.66	37.15	45.15	96.6	46.98	87.53	36.37	42.22	33.72	96.26	
BCR(O)	22	15	48.91	144.97	41.82	148.74	91.31	133.13	60.03	231.16	100.21	141.10	135.46
FMS(O)	20	15	67.34	111.08	64.01	112.30	71.93	120.70	38.89	78.17	47.08	148.17	147.22
BCL-3(O)	8	8	67.66	73.05	166.48	61.41	73.02	147.51	71.24	219.16			
FOS(O)	3	3	13.01	74.63	17.06								
HST-1(O)	2	2	45.96	132.73									
INT-2(O)	2	2	27.69	191.89									
LCK(O)	11	11	46.08	73.61	86.83	71.10	84.16	75.41	60.89	91.18	37.55	29.26	40.35
MYC(O)	2	2	9.20	56.08									
L-MYC(O)	2	2	39.76	81.53									
N-MYC(O)	2	2	11.16	71.24									
MAX(O)	1	1	36.28										
PIM1(O)	5	5	18.1	78.28	74.98	88.41	37.01						
KIT(O)	20	15	9.41	33.80	15.91	25.39	17.75	55.72	19.88	16.36	10.23	9.87	23.95
RAFA1	15	15	20.37	69.19	36.24	51.01	47.33	33.21	34.60	68.25	29.27	54.91	48.19
FPS(O)	18	15	35.54	83.33	78.68	59.44	74.13	66.73	60.31	86.15	79.85	107.85	116.19
WNT-1(O)	3	3	23.28	115.26	96.70								
K-RAS2(O)	4	4	18.87	32.27	15.53	31.64							
MLH1(TS)	18	15	11.15	29.71	9.16	23.94	12.96	34.83	10.46	28.77	20.58	29.50	30.25
MTS1(TS)	3	3	42.77	131.26	49.18								
RB(TS)	26	15	11.71	23.38	13.20	16.98	13.85	35.74	18.01	9.19	21.83	34.09	9.71
NF1(TS)	56	15	22.20	22.24	11.45	39.09	12.96	17.59	11.13	10.17	17.13	17.49	17.46
G6PD(NR)	12	12	12.92	108.50	66.24	117.07	94.67	97.99	61.70	146.62	64.08	52.00	54.42
PGK(NR)	10	10	13.37	27.74	38.02	37.56	32.20	38.87	18.39	28.48	27.09	47.37	
ADOL(NR)	8	8	106.52	75.91	32.15	46.29	30.70	47.30	32.50	39.65			
GTP(NR)	8	8	24.28	52.12	36.08	62.00	19.50	54.17	19.87	54.42			
PFK(NR)	21	15	17.20	55.42	53.34	112.29	62.74	81.66	120.00	77.67	34.79	107.84	60.25
TPI(NR)	6	6	24.52	57.34	14.68	72.53	51.39	70.67					
GCK(NR)	9	9	58.22	107.13	102.01	81.17	70.17	88.17	56.29	136.99	73.35		
GAPDH(NR)	8	8	28.43	129.01	41.90	37.78	36.87	58.32	47.54	42.90			
SDHB(NR)	7	7	20.49	31.56	59.52	50.69	38.14	71.55	34.25				
HKII(NR)	17	15	16.00	41.70	41.78	66.23	61.01	73.82	36.48	67.54	42.11	28.03	76.14

Table 6.20 (Continued): MSH2(TS) Gene (15 Introns) Versus Contrast Genes (Gray Cells Reject Ho at $\alpha= 0.01$ Level)

Contrast Gene	Intron 12	Intron 13	Intron 14	Intron 15	Mean	SD
ABL(O)					59.16	25.77
BCR(O)	121.61	76.92	26.75	75.83	105.20	53.36
FMS(O)	42.27	88.59	25.90	80.00	82.91	38.16
BCL-3(O)					109.94	59.64
FOS(O)					34.90	34.47
HST-1(O)					89.35	61.36
INT-2(O)					109.79	116.10
LCK(O)					63.31	21.75
MYC(O)					32.64	33.15
L-MYC(O)					60.65	29.54
N-MYC(O)					41.20	42.48
MAX(O)					36.28	NA
PIM1(O)					59.36	30.20
KIT(O)	15.01	7.72	11.17	18.44	19.37	12.27
RAFA1(O)	61.51	73.62	17.51	97.56	49.52	21.97
FPS(O)	76.03	91.90	25.83	71.59	74.24	23.60
WNT-1(O)					78.41	48.64
K-RAS2(O)					24.58	8.63
MLH1(TS)	28.31	23.38	21.18	15.95	22.01	8.35
MTS1(TS)					74.40	49.34
RB(TS)	16.94	16.57	22.11	29.20	19.50	8.26
NF1(TS)	14.41	16.22	15.28	16.08	17.39	6.97
G6PD(NR)	49.78				77.17	36.55
PGK(NR)					30.91	10.14
ADOL(NR)					51.38	26.66
GTP(NR)					40.30	17.43
PFK(NR)	45.51	53.02	15.95	45.31	62.87	31.89
TPI(NR)					48.52	23.97
GCK(NR)					85.94	25.97
GAPDH(NR)					52.84	31.97
SDHB(NR)					43.74	17.71
HKII(NR)	15.95	26.84	22.98	27.89	42.97	20.93

Table C.21: MTS1(TS) Gene (3 Introns) Versus Contrast Genes (Gray Cells Reject Ho at α = 0.01 Level)

Contrast Gene	# Introns in Contrast Gene	# Introns Compared	Intron 1	Intron 2	Intron 3	Mean	SD
ABL(O)	10	3	74.75	90.99	26.85	64.20	33.34
BCR(O)	22	3	40.37	26.35	7.91	24.88	16.28
FMS(O)	21	3	54.48	45.53	14.27	38.09	21.11
BCL-3(O)	8	3	34.18	67.49	106.29	69.32	36.09
FOS(O)	3	3	41.86	25.00	14.20	27.02	13.94
HST-1(O)	2	2	81.37	13.21		47.29	48.20
INT-2(O)	2	2	68.03	66.12		67.07	1.35
LCK(O)	11	3	32.25	18.13	54.42	34.93	18.29
MYC(O)	2	2	65.72	69.34		67.53	2.56
L-MYC(O)	2	2	73.48	58.28		65.88	10.75
N-MYC(O)	2	2	45.34	61.30		53.32	11.29
MAX(O)	1	1	30.41			30.41	NA
PIM1(O)	5	3	43.42	50.43	47.65	47.17	3.53
KIT(O)	20	3	32.49	51.28	23.53	35.77	14.17
RAFA1(O)	15	3	44.03	16.38	24.33	28.25	14.23
FPS(O)	18	3	50.92	25.06	19.16	31.71	16.89
WNT-1(O)	3	3	57.77	36.46	64.80	53.01	14.76
K-RAS2(O)	4	3	58.07	107.75	64.93	76.92	26.92
MLH1(TS)	18	3	50.84	66.85	27.06	48.25	20.02
MSH2(TS)	15	3	42.77	131.26	49.18	74.40	49.34
RB(TS)	26	3	49.32	134.04	67.37	83.57	44.63
NF1(TS)	56	3	42.70	60.67	54.00	52.46	9.08
G6PD(NR)	12	3	37.48	45.58	30.13	37.73	7.73
PGK(NR)	10	3	49.56	69.55	29.62	49.57	19.97
ADOL(NR)	8	3	59.67	57.58	27.92	48.39	17.76
GTP(NR)	8	3	20.39	54.57	18.50	31.15	20.30
PFK(NR)	21	3	17.69	5.85	25.26	16.27	9.78
TPI(NR)	6	3	56.76	13.31	15.00	28.35	24.61
GCK(NR)	9	3	50.61	15.79	56.84	41.08	22.13
GAPDH(NR)	8	3	44.69	28.99	12.93	28.87	15.88
SDHB(NR)	7	3	29.15	61.49	24.23	38.29	20.24
HKII(NR)	17	3	34.97	25.57	13.24	24.60	10.90

Table C.22: RB(TS) Gene (26 Introns) Versus Contrast Genes (Gray Cells Reject Ho at α = 0.01 Level)

Contrast Gene	# Introns in Contrast Gene	# Introns Compared	Intron 1	Intron 2	Intron 3	Intron 4	Intron 5	Intron 6	Intron 7	Intron 8	Intron 9	Intron 10	Intron 11	Intron 12	Intron 13	Intron 14	Intron 15
ABL(O)	10	10	66.01	12.64	61.63	119.57	68.41	90.73	99.96	35.04	49.15	139.72					
BCR(O)	22	22	74.50	187.22	44.82	218.95	145.29	174.65	94.43	199.95	134.15	179.95	129.08	203.48	93.75		
FMS(O)	21	21	58.01	127.35	71.11	178.72	80.00	154.20	67.59	81.17	61.21	194.19	137.95	75.51	93.53	47.72	91.03
BCL-3(O)	8	8	93.70	63.93	189.78	71.80	112.84	180.83	90.56	189.34							
FOS(O)	3	3	41.63	60.10	13.77												
HST-1(O)	2	2	97.93	137.39													
INT-2(O)	2	2	109.26	240.71													
LCK(O)	11	11	58.14	61.06	92.63	85.70	122.32	93.71	62.55	78.48	41.08	48.51	29.60				
MYC(O)	2	2	34.71	51.96													
L-MYC(O)	2	2	78.95	74.69													
N-MYC(O)	2	2	43.75	63.88													
MAX(O)	1	1	67.68														
PIM1(O)	5	5	32.98	81.14	68.76	100.67	39.27										
KIT(O)	20	20	13.38	14.61	17.70	13.69	12.40	74.87	23.66	13.37	21.06	18.57	20.72	19.04	13.78	26.18	47.42
RAFA1(O)	15	15	42.50	42.34	62.20	59.05	54.83	35.05	38.69	54.00	47.89	70.02	38.37	122.59	77.77	54.46	93.43
FPS(O)	18	18	85.54	76.72	97.04	69.37	93.64	66.80	82.99	72.39	84.34	150.42	98.54	122.01	93.15	36.55	72.40
WNT-1(O)	3	3	59.12	133.86	104.71												
K-RAS2(O)	4	4	22.56	20.30	13.00	16.01											
MLH1(TS)	18	18	7.43	11.05	10.96	17.76	8.85	28.38	13.46	40.08	31.89	49.42	20.22	37.01	18.27	20.54	35.75
MSH2(TS)	15	15	11.71	23.38	13.20	16.98	13.85	35.74	18.01	9.19	21.83	34.09	9.71	16.94	16.57	22.11	29.20
MTS1(TS)	3	3	49.32	134.04	67.37												
NF1(TS)	56	26	22.84	23.65	14.41	26.44	8.15	18.64	11.54	14.43	6.45	14.59	16.43	7.50	20.70	39.94	22.60
G6PD(NR)	12	12	29.02	135.05	67.54	164.91	139.66	107.61	80.94	134.95	71.70	76.96	45.39	82.86			
PGK(NR)	10	10	17.52	13.18	46.21	26.10	23.85	30.91	24.30	27.49	31.29	47.91					
ADOL(NR)	8	8	102.80	60.63	34.25	43.36	23.73	38.61	52.34	32.50							
GTP(NR)	8	8	34.39	25.00	28.45	71.90	16.86	62.96	41.76	39.61							
PFK(NR)	21	21	28.78	34.50	48.29	145.45	77.27	108.46	222.64	80.43	57.73	149.17	57.18	64.66	46.03	16.35	57.73
TPI(NR)	6	6	76.99	44.88	19.49	87.62	57.05	72.30									
GCK(NR)	9	9	119.70	98.55	122.85	117.94	74.96	103.11	76.57	135.24	93.10						
GAPDH(NR)	8	8	54.00	159.96	40.23	55.15	47.08	62.58	59.98	42.27							
SDHB(NR)	7	7	22.96	13.64	57.90	45.82	44.13	60.03	42.02								
HKII(NR)	17	17	32.56	27.31	44.86	65.90	79.03	106.36	58.03	66.77	59.35	46.69	70.51	26.93	29.48	25.20	52.76

Table C.22 (Continued): RB(TS) Gene (26 Introns) Versus Contrast Genes (Gray Cells Reject Ho at α = 0.01 Level)

Contrast Gene	Intron 16	Intron 17	Intron 18	Intron 19	Intron 20	Intron 21	Intron 22	Intron 23	Intron 24	Intron 25	Intron 26	Mean	SD
ABL(O)												74.29	38.73
BCR(O)	137.83	66.11	163.11	92.49	146.12	189.88	37.38					128.72	57.16
FMS(O)	51.94	76.80	51.12	60.08	39.98	77.12						89.35	43.66
BCL-3(O)												124.10	53.88
FOS(O)												38.50	23.32
HST-1(O)												117.66	27.90
INT-2(O)												174.99	92.95
LCK(O)												70.34	27.08
MYC(O)												43.34	12.20
L-MYC(O)												76.82	3.01
N-MYC(O)												53.82	14.23
MAX(O)												67.68	NA
PIM1(O)												64.57	28.43
KIT(O)	15.22	11.54	23.73	16.53	28.16							22.28	14.80
RAFA1(O)												59.54	23.69
FPS(O)	86.63	84.01	118.64									88.40	24.69
WNT-1(O)												99.23	37.67
K-RAS2(O)	17.34	12.02	47.30									17.97	4.28
MLH1(TS)												23.76	13.45
MSH2(TS)												19.50	8.26
MTS1(TS)												83.58	44.62
NF1(TS)	8.03	10.726	20.947	17.746	18.305	37.96	10.956	11.372	12.509	16.049	7.59	16.94	8.54
G6PD(NR)												94.71	41.58
PGK(NR)												28.88	11.07
ADOL(NR)												48.53	24.81
GTP(NR)												40.12	18.80
PFK(NR)	29.93	74.414	47.196	39.471	48.227	150.85						75.46	51.84
TPI(NR)												59.72	24.82
GCK(NR)												104.67	20.96
GAPDH(NR)												65.16	39.13
SDHB(NR)												40.93	17.11
HKII(NR)	35.83	31.447										50.53	22.50

133

Table C.23: NF1(TS) Gene (56 Introns) Versus Contrast Genes (Gray Cells Reject Ho at α = 0.01 Level)

Contrast Gene	# Introns in Contrast Gene	# Introns Compared	Intron 1	Intron 2	Intron 3	Intron 4	Intron 5	Intron 6	Intron 7	Intron 8	Intron 9	Intron 10	Intron 11	Intron 12	Intron 13	Intron 14	Intron 15
ABL(O)	10	10	24.74	19.93	61.52	21.15	76.12	48.06	35.54	42.57	39.79	56.09					
BCR(O)	22	22	38.60	49.84	40.08	55.58	170.32	84.30	62.32	181.95	110.47	71.78	197.99	81.96	54.41	1833.?	?.04
FMS(O)	21	21	34.83	39.18	67.66	64.68	102.82	65.85	32.35	65.34	50.75	77.99	277.17	53.39	70.07	110.23	85.73
BCL-3(O)	8	8	62.87	27.82	183.86	39.45	115.75	139.41	66.97	160.33							
FOS(O)	3	3	43.44	25.79	14.98												
HST-1(O)	2	2	67.50	60.70													
INT-2(O)	2	2	64.05	74.01													
LCK(O)	11	11	49.78	37.08	97.08	47.14	122.52	65.14	53.50	73.12	34.70	34.39	46.89				
MYC(O)	2	2	34.39	22.87													
L-MYC(O)	2	2	52.80	31.73													
N-MYC(O)	2	2	37.25	27.16													
MAX(O)	1	1	49.91														
PIM1(O)	5	5	39.85	56.26	68.99	23.08	46.26										
KIT(O)	20	20	22.27	14.16	22.82	18.44	6.67	41.84	14.02	14.44	17.04	11.49	33.49	17.52	11.00	11.55	27.61
RAFA1(O)	15	15	39.90	32.82	55.10	29.57	57.75	23.21	29.67	64.56	39.06	35.47	57.25	59.76	76.42	104.20	91.09
FPS(O)	18	18	63.64	46.39	96.28	23.87	113.64	59.55	57.55	70.14	78.63	62.59	137.69	69.29	54.74	81.50	79.03
WNT-1(O)	3	3	52.20	60.46	111.95												
K-RAS2(O)	4	4	30.94	15.10	11.05	24.52											
MLH1(TS)	18	18	20.35	13.33	12.46	22.82	9.36	11.86	13.36	30.22	25.03	33.88	34.63	25.41	21.88	40.36	22.72
MSH2(TS)	15	15	22.20	22.24	11.45	39.09	12.96	17.59	11.13	10.17	17.13	17.49	17.46	14.41	16.22	15.28	16.08
MTS1(TS)	3	3	42.70	60.67	54.00												
RB(TS)	26	26	22.84	23.65	14.41	26.44	8.15	18.64	11.54	14.43	6.45	14.59	16.43	7.50	20.70	39.94	22.60
G6PD(NR)	12	12	35.99	37.60	63.86	66.76	156.46	81.35	54.68	127.86	63.17	54.37	62.23	54.55			
PGK(NR)	10	10	20.43	9.32	47.95	8.36	27.69	20.89	14.78	39.14	26.60	30.05					
ADOL(NR)	8	8	37.40	34.82	41.87	14.61	32.89	22.61	25.55	35.78							
GTP(NR)	8	8	37.04	23.98	32.48	27.57	18.14	46.07	22.11	47.15							
PFK(NR)	21	21	33.55	29.30	57.03	70.74	80.61	72.00	132.50	75.01	44.72	77.71	75.26	45.80	46.86	32.03	44.97
TPI(NR)	6	6	57.71	25.93	21.01	32.87	78.82	55.62									
GCK(NR)	9	9	62.22	51.93	113.10	50.66	80.59	74.07	49.81	106.26	79.66						
GAPDH(NR)	8	8	49.64	58.23	36.68	20.66	48.98	55.39	44.18	38.41							
SDHB(NR)	7	7	25.92	14.75	66.06	13.97	50.26	43.24	25.81								
HKII(NR)	17	17	31.56	20.18	49.67	26.60	83.04	64.54	35.30	66.09	49.12	27.43	88.06	24.24	35.76	42.22	40.55

Table C.23 (Continued): NF1(TS) Gene (56 Introns) Versus Contrast Genes (Gray Cells Reject Ho at α = 0.01 Level)

Contrast Gene	Intron 16	Intron 17	Intron 18	Intron 19	Intron 20	Intron 21	Intron 22	Intron 23	Intron 24	Intron 25	Intron 26	Mean	SD
ABL(O)												42.55	18.39
BCR(O)	53.40	54.77	110.27	54.11	68.08	102.25	82.21					87.99	45.96
FMS(O)	39.36	56.38	34.03	37.09	25.56	31.69						68.10	53.77
BCL-3(O)												99.56	58.33
FOS(O)												28.07	14.37
HST-1(O)												64.10	4.81
INT-2(O)												69.03	7.04
LCK(O)												60.12	27.91
MYC(O)												28.63	8.15
L-MYC(O)												42.27	14.90
N-MYC(O)												32.21	7.14
MAX(O)												49.91	NA
PIM1(O)												46.89	17.27
KIT(O)	11.84	6.62	14.61	10.42	18.91							17.34	8.82
RAFA1(O)												53.05	23.76
FPS(O)	60.10	66.10	88.77									72.75	25.52
WNT-1(O)												74.87	32.38
K-RAS2(O)												20.40	9.01
MLH1(TS)	14.85	14.00	41.17									22.65	10.03
MSH2(TS)												17.39	6.97
MTS1(TS)												52.46	9.08
RB(TS)	8.03	10.73	20.95	17.75	18.30	37.96	10.96	11.37	12.51	16.05	7.59	16.94	8.54
G6PD(NR)												71.57	35.68
PGK(NR)												24.52	12.58
ADOL(NR)												30.69	9.01
GTP(NR)												31.82	10.87
PFK(NR)	31.70	56.23	37.25	30.14	33.27	88.84						56.93	26.03
TPI(NR)												44.00	20.11
GCK(NR)												74.26	23.45
GAPDH(NR)												44.02	12.06
SDHB(NR)												34.29	19.50
HKII(NR)	29.45	17.00										42.98	21.21

Table C.24: G6PD(NR) Gene (12 Introns) Versus Contrast Genes (Gray Cells Reject Ho at α = 0.01 Level))

Contrast Gene	# Introns in Contrast Gene	# Introns Compared	Intron 1	Intron 2	Intron 3	Intron 4	Intron 5	Intron 6	Intron 7	Intron 8	Intron 9	Intron 10	Intron 11	Intron 12	Mean	SD
ABL(O)	10	10	86.96	80.03	38.39	79.58	46.95	46.05	26.70	101.33	16.61	19.65			54.23	30.44
BCR(O)	22	12	40.84	52.48	21.53	43.71	44.15	25.29	11.56	22.28	14.55	12.79	14.04	16.72	26.66	14.58
FMS(O)	21	12	63.92	60.19	19.08	56.69	44.39	53.97	29.34	21.01	7.68	13.12	25.04	11.16	33.80	20.77
BCL-3(O)	8	8	58.22	36.28	49.76	21.18	24.22	34.84	23.87	75.45					40.48	19.22
FOS(O)	3	3	14.67	10.10	30.65										18.47	10.79
HST-1(O)	2	2	56.58	42.73											49.66	9.79
INT-2(O)	2	2	32.55	243.56											138.06	149.20
LCK(O)	11	11	33.06	18.75	23.82	29.07	42.27	14.08	14.19	33.68	8.06	8.13	31.03		23.29	11.43
MYC(O)	2	2	23.41	45.22											34.32	15.42
L-MYC(O)	2	2	49.69	62.01											55.85	8.71
N-MYC(O)	2	2	14.16	46.75											30.45	23.04
MAX(O)	1	1	27.73												27.73	NA
PIM1(O)	5	5	22.95	74.84	13.72	74.90	86.23								54.52	33.52
KIT(O)	20	12	6.20	34.02	53.21	100.63	118.58	40.65	21.45	126.99	62.54	35.29	36.43	39.62	56.30	38.69
RAFA1(O)	15	12	7.40	18.17	38.18	11.98	14.36	92.29	4.54	17.90	9.86	29.82	13.18	14.25	22.66	23.84
FPS(O)	18	12	40.29	40.98	24.04	4.94	53.11	9.23	17.84	31.89	15.97	15.95	15.25	13.33	23.57	14.77
WNT-1(O)	3	3	13.89	76.72	33.98										41.53	32.09
K-RAS2(O)	4	4	37.90	84.42	66.83	178.38									91.88	60.77
MLH1(TS)	18	12	21.40	38.85	47.92	72.09	86.08	68.36	42.33	107.01	44.86	27.44	40.62	37.25	52.85	25.35
MSH2(TS)	15	12	12.92	108.50	66.24	117.07	94.67	97.99	61.70	146.62	64.08	52.00	54.42	49.78	77.17	36.55
MTS1(TS)	3	3	37.48	45.58	30.13										37.73	7.73
RB(TS)	26	12	29.02	135.05	67.54	164.91	139.66	107.61	80.94	134.95	71.70	76.96	45.39	82.86	94.71	41.58
NF1(TS)	56	12	35.99	37.596	63.864	66.761	156.46	81.349	54.676	127.86	63.175	54.374	62.229	54.551	71.57	35.68
PGK(NR)	10	10	15.14	44.82	34.69	67.70	77.18	47.35	43.54	60.00	36.46	27.11			45.40	18.78
ADOL(NR)	8	8	108.92	50.68	52.79	98.41	117.54	44.46	32.04	120.70					78.19	36.60
GTP(NR)	8	8	14.07	27.13	22.02	35.56	88.36	16.06	29.35	46.73					34.91	24.03
PFK(NR)	21	12	17.31	10.54	17.53	23.37	10.81	22.21	34.64	18.27	10.66	6.15	10.13	8.91	15.88	8.07
TPI(NR)	6	6	26.71	13.01	34.75	12.55	35.20	11.26							22.25	11.35
GCK(NR)	9	9	37.96	23.62	21.98	26.12	22.58	6.80	13.03	43.86	29.11				25.01	11.36
GAPDH(NR)	8	8	46.84	89.38	8.21	19.85	8.04	12.77	6.91	25.53					27.19	28.43
SDHB(NR)	7	7	23.79	37.15	34.97	59.81	34.74	36.21	19.20						35.12	12.89
HKII(NR)	17	12	24.93	15.71	16.09	24.58	46.26	22.96	19.66	24.82	8.65	21.38	15.56	21.13	21.81	9.11

Table C.25: PGK(NR) Gene (10 Introns) Versus Contrast Genes (Gray Cells Reject Ho at α = 0.01 Level)

Contrast Gene	# Introns in Contrast Gene	# Introns Compared	Intron 1	Intron 2	Intron 3	Intron 4	Intron 5	Intron 6	Intron 7	Intron 8	Intron 9	Intron 10	Mean	SD
ABL(O)	10	10	25.64	9.07	29.02	18.50	34.73	21.43	31.14	17.91	22.43	27.47	23.73	7.49
BCR(O)	22	10	23.54	64.43	25.80	48.28	54.22	40.05	42.34	54.57	45.31	25.38	42.39	13.91
FMS(O)	21	10	37.75	44.80	36.05	51.17	28.66	30.96	28.08	40.49	29.59	40.65	36.82	7.67
BCL-3(O)	8	8	69.28	27.46	87.91	44.61	52.15	80.67	45.97	111.31			64.92	27.48
FOS(O)	3	3	31.66	23.30	19.03								24.66	6.42
HST-1(O)	2	2	81.82	71.31									76.57	7.43
INT-2(O)	2	2	69.54	94.78									82.16	17.84
LCK(O)	11	10	46.68	36.46	41.99	45.32	71.97	37.14	46.28	38.98	34.41	14.31	41.35	14.27
MYC(O)	2	2	26.54	20.36									23.45	4.37
L-MYC(O)	2	2	65.06	26.11									45.59	27.54
N-MYC(O)	2	2	26.32	20.50									23.41	4.12
MAX(O)	1	1	39.30										39.30	NA
PIM1(O)	5	5	30.47	62.50	51.48	36.83	19.94						40.24	16.89
KIT(O)	20	10	11.29	8.08	26.59	15.08	16.77	25.53	17.88	26.36	25.78	19.66	19.30	6.67
RAFA1(O)	15	10	18.20	28.64	33.34	24.59	25.23	18.09	24.51	42.59	27.64	19.64	26.25	7.50
FPS(O)	18	10	63.25	43.64	32.9	31.93	29.35	34.20	48.37	58.99	31.08	40.66	41.44	12.05
WNT-1(O)	3	3	37.11	68.43	48.38								51.31	15.86
K-RAS2(O)	4	4	35.64	13.70	46.70	20.43							29.12	14.89
MLH1(TS)	18	10	16.68	7.56	29.36	21.88	20.40	7.37	11.17	15.42	34.32	24.80	18.90	9.01
MSH2(TS)	15	10	13.37	27.74	38.02	37.56	32.20	38.87	18.39	28.48	27.09	47.37	30.91	10.14
MTS1(TS)	3	3	49.56	69.55	29.62								49.57	19.97
RB(TS)	26	10	17.52	13.18	46.21	26.10	23.85	30.91	24.30	27.49	31.29	47.91	28.88	11.07
NF1(TS)	56	10	20.43	9.32	47.95	8.36	27.69	20.89	14.78	39.14	26.60	30.05	24.52	12.58
G6PD(NR)	12	10	15.14	44.82	34.69	67.70	77.18	47.35	43.54	60.00	36.46	27.11	45.40	18.78
ADOL(NR)	8	8	53.79	26.92	33.46	8.17	10.23	8.82	30.58	22.40			24.30	15.59
GTP(NR)	8	8	21.93	11.89	23.03	17.19	16.08	24.83	18.29	15.11			18.54	4.39
PFK(NR)	21	10	21.72	23.05	32.03	77.66	40.14	40.16	81.17	45.67	26.89	44.92	43.34	20.86
TPI(NR)	6	6	53.96	25.72	24.08	33.75	25.51	29.14					32.03	11.29
GCK(NR)	9	9	54.54	50.87	58.22	47.88	43.62	38.08	29.33	83.23	42.94		49.86	15.24
GAPDH(NR)	8	8	60.50	66.35	31.00	24.01	25.21	34.68	35.98	22.11			37.48	16.84
SDHB(NR)	7	7	14.77	10.34	31.21	12.65	21.49	24.20	32.33				21.00	8.81
HKII(NR)	17	10	25.32	11.69	21.78	21.21	26.73	47.07	27.90	28.26	34.15	24.78	26.89	9.18

Table C.26: ADOL(NR) Gene (8 Introns) Versus Contrast Genes (Gray Cells Reject Ho at α = 0.01 Level)

Contrast Gene	# Introns in Contrast Gene	# Introns Compared	Intron 1	Intron 2	Intron 3	Intron 4	Intron 5	Intron 6	Intron 7	Intron 8	Mean	SD
ABL(O)	10	8	69.54	34.27	57.03	39.79	49.89	12.69	35.10	70.33	46.08	19.59
BCR(O)	22	8	226.34	65.30	30.21	109.66	106.37	27.06	42.34	458.35	133.20	146.59
FMS(O)	21	8	9.85	35.51	40.14	87.12	43.03	25.00	18.98	62.01	40.21	24.81
BCL-3(O)	8	8	143.01	27.39	197.38	56.11	99.32	85.87	74.88	337.32	127.66	99.69
FOS(O)	3	3	214.12	17.03	12.40						81.18	115.15
HST-1(O)	2	2	357.69	53.24							205.47	215.28
INT-2(O)	2	2	533.91	111.08							322.50	298.99
LCK(O)	11	8	49.97	34.55	92.18	73.06	126.52	37.47	53.09	78.13	68.12	31.02
MYC(O)	2	2	311.7	37.21							174.46	194.09
L-MYC(O)	2	2	288.76	29.58							159.17	183.27
N-MYC(O)	2	2	182.91	31.78							107.34	106.86
MAX(O)	1	1	72.53								72.53	NA
PIM1(O)	5	5	122.17	101.20	75.60	67.66	16.94				76.71	39.76
KIT(O)	20	8	72.98	18.94	12.55	29.47	20.70	20.60	17.41	44.66	29.66	20.08
RAFA1(O)	15	8	159.82	26.11	44.35	31.98	33.05	28.65	23.40	57.20	50.57	45.49
FPS(O)	18	8	202.28	44.98	69.58	37.07	58.32	30.13	55.95	88.89	73.40	55.29
WNT-1(O)	3	3	169.16	84.46	114.08						122.57	42.98
K-RAS2(O)	4	4	168.1	57.64	46.29	22.52					73.64	64.65
MLH1(TS)	18	8	99.45	26.75	13.21	32.05	16.50	11.47	29.22	38.95	33.45	28.36
MSH2(TS)	15	8	106.52	75.91	32.15	46.29	30.70	47.30	32.50	39.65	51.38	26.66
MTS1(TS)	3	3	59.67	57.58	27.92						48.39	17.76
RB(TS)	26	8	102.8	60.63	34.25	43.36	23.73	38.61	52.34	32.50	48.53	24.81
NF1(TS)	56	8	37.4	34.82	41.87	14.61	32.89	22.61	25.55	35.78	30.69	9.01
G6PD(NR)	12	8	108.92	50.68	52.79	98.41	117.54	44.46	32.04	120.70	78.19	36.60
PGK(NR)	10	8	53.79	26.92	33.46	8.17	10.23	8.82	30.58	22.40	24.30	15.59
GTP(NR)	8	8	45.20	27.81	19.83	27.56	12.50	19.64	19.98	29.86	25.30	9.87
PFK(NR)	21	8	34.72	21.72	38.72	95.87	46.24	37.35	122.31	58.81	56.97	34.55
TPI(NR)	6	6	320.17	22.70	12.67	42.06	29.97	28.88			76.07	119.97
GCK(NR)	9	8	248.10	38.95	114.80	51.29	52.34	32.50	47.10	172.25	94.67	78.26
GAPDH(NR)	8	8	273.30	87.16	30.97	32.61	26.91	28.85	37.76	28.00	68.19	85.25
SDHB(NR)	7	7	41.86	30.15	34.96	21.23	18.58	11.30	22.25		25.76	10.46
HKII(NR)	17	8	45.58	17.81	21.33	33.37	47.50	39.53	41.59	49.01	36.97	11.85

Table C.27: GTP(NR) Gene (8 Introns) Versus Contrast Genes (Gray Cells Reject Ho at $\alpha = 0.01$ Level)

Contrast Gene	# Introns in Contrast Gene	# Introns Compared	Intron 1	Intron 2	Intron 3	Intron 4	Intron 5	Intron 6	Intron 7	Intron 8	Mean	SD
ABL(O)	10	8	49.50	15.81	15.65	16.39	25.30	18.31	15.96	38.71	24.46	12.86
BCR(O)	22	8	15.73	47.53	10.36	18.77	72.07	8.54	31.60	46.03	31.33	22.35
FMS(O)	21	8	38.05	32.95	16.48	21.75	51.05	21.12	14.80	38.41	29.33	12.79
BCL-3(O)	8	8	45.91	15.29	52.37	23.19	68.59	22.44	48.86	69.64	43.29	20.97
FOS(O)	3	3	32.95	19.22	15.77						22.65	9.09
HST-1(O)	2	2	84.29	58.17							71.23	18.47
INT-2(O)	2	2	60.46	77.49							68.98	12.04
LCK(O)	11	8	30.96	29.14	24.21	24.51	81.30	15.06	35.47	29.51	33.77	20.13
MYC(O)	2	2	43.98	22.11							33.05	15.46
L-MYC(O)	2	2	71	20.85							45.93	35.46
N-MYC(O)	2	2	25.32	13.80							19.56	8.15
MAX(O)	1	1	15.16								15.16	NA
PIM1(O)	5	5	36.99	59.17	22.40	33.79	18.91				34.25	15.85
KIT(O)	20	8	18.19	16.03	25.03	43.50	14.37	17.84	7.92	43.75	23.33	13.38
RAFA1(O)	15	8	18.18	16.36	23.21	18.48	35.18	45.94	14.13	37.21	26.09	11.75
FPS(O)	18	8	46.91	42.99	14.40	18.24	45.11	8.92	42.94	49.48	33.62	16.69
WNT-1(O)	3	3	39.26	60.48	18.84						39.53	20.82
K-RAS2(O)	4	4	53.35	21.64	31.15	63.15					42.32	19.22
MLH1(TS)	18	8	35.39	11.34	22.39	44.99	10.45	30.83	14.39	37.47	25.91	13.14
MSH2(TS)	15	8	24.28	52.12	36.08	62.00	19.50	54.17	19.87	54.42	40.30	17.43
MTS1(TS)	3	3	20.39	54.57	18.50						31.15	20.30
RB(TS)	26	8	34.39	25.00	28.45	71.90	16.86	62.96	41.76	39.61	40.12	18.80
NF1(TS)	56	8	37.04	23.98	32.48	27.57	18.14	46.07	22.11	47.15	31.82	10.87
G6PD(NR)	12	8	14.07	27.13	22.02	35.56	88.36	16.06	29.35	46.73	34.91	24.03
PGK(NR)	10	8	21.93	11.89	23.03	17.19	16.08	24.83	18.29	15.11	18.54	4.39
ADOL(NR)	8	8	45.20	27.81	19.83	27.56	12.50	19.64	19.98	29.86	25.30	9.87
PFK(NR)	21	8	18.18	14.29	21.17	48.22	42.56	9.83	104.04	46.09	38.05	30.67
TPI(NR)	6	6	44.29	22.15	13.61	14.68	39.41	8.77			23.82	14.69
GCK(NR)	9	8	31.36	38.60	32.82	23.01	54.94	10.15	28.23	70.03	36.14	18.72
GAPDH(NR)	8	8	59.50	64.42	11.61	18.58	23.29	8.54	24.43	18.64	28.63	21.30
SDHB(NR)	7	7	15.91	16.91	10.40	23.77	16.04	14.67	13.93		15.95	4.06
HKII(NR)	17	8	26.28	11.90	15.00	23.43	46.25	20.36	22.84	34.77	25.11	11.01

Table C.28: PFK(NR) Gene (21 Introns) Versus Contrast Genes (Gray Cells Reject Ho at α= 0.01 Level)

Contrast Gene	# Introns in Contrast Gene	# Introns Compared	Intron 1	Intron 2	Intron 3	Intron 4	Intron 5	Intron 6	Intron 7	Intron 8	Intron 9	Intron 10	Intron 11
ABL(O)	10	10	34.3	26.72	28.2	69.98	23.18	37.41	140.16	50.24	9.03	43.79	
BCR(O)	22	21	17.22	7.92	18.31	39.86	13.59	24.02	40.08	22.11	6.24	23.34	10.21
FMS(O)	21	21	33.51	13.08	13.44	38.14	20.39	38.30	203.66	15.95	8.44	25.18	26.82
BCL-3(O)	8	8	14.79	17.71	31.34	19.49	20.16	37.92	29.52	73.15			
FOS(O)	3	3	12.96	12.13	25.47								
HST-1(O)	2	2	21.91	11.30									
INT-2(O)	2	2	13.95	11.99									
LCK(O)	11	11	19.97	9.40	15.60	17.42	32.33	26.30	21.41	27.07	9.14	5.31	32.18
MYC(O)	2	2	18.41	16.57									
L-MYC(O)	2	2	18.65	14.17									
N-MYC(O)	2	2	12.20	13.26									
MAX(O)	1	1	10.88										
PIM1(O)	5	5	13.62	24.17	19.57	51.36	35.75						
KIT(O)	20	20	17.62	20.33	36.64	103.35	63.19	28.28	75.10	67.95	42.78	43.50	41.54
RAFA1(O)	15	15	17.35	11.78	30.54	16.48	10.61	85.44	9.36	18.90	5.03	65.21	12.22
FPS(O)	18	18	8.47	8.19	16.74	11.98	20.32	11.65	15.33	21.16	24.82	10.78	21.68
WNT-1(O)	3	3	14.82	13.89	23.08								
K-RAS2(O)	4	4	23.69	29.20	62.56	153.60							
MLH1(TS)	18	18	27.68	27.45	41.09	79.87	50.37	56.14	88.76	57.67	27.51	51.44	47.98
MSH2(TS)	15	15	17.20	55.42	53.34	112.29	62.74	81.66	120.00	77.67	34.79	107.84	60.25
MTS1(TS)	3	3	17.69	5.85	25.26								
RB(TS)	26	21	28.78	34.50	48.29	145.45	77.27	108.46	222.64	80.43	57.73	149.17	57.18
NF1(TS)	56	21	33.55	29.30	57.03	70.74	80.61	72.00	132.50	75.01	44.72	77.71	75.26
G6PD(NR)	12	12	17.31	10.54	17.53	23.37	10.81	22.21	34.64	18.27	10.66	6.15	10.13
PGK(NR)	10	10	21.72	23.05	32.03	77.66	40.14	40.16	81.17	45.67	26.89	44.92	
ADOL(NR)	8	8	34.72	21.72	38.72	95.87	46.24	37.35	122.31	58.81			
GTP(NR)	8	8	18.18	14.29	21.17	48.22	42.56	9.83	104.04	46.09			
TPI(NR)	6	6	14.00	13.39	27.18	28.99	22.38	13.37					
GCK(NR)	9	9	13.02	10.34	20.72	26.82	18.87	15.72	24.88	21.04	14.08		
GAPDH(NR)	8	8	16.23	8.02	12.74	32.84	6.14	12.47	22.98	16.70			
SDHB(NR)	7	7	19.06	23.98	23.74	63.49	16.31	26.36	82.74				
HKII(NR)	17	17	16.83	11.77	12.99	34.47	17.85	11.77	24.16	23.24	3.79	29.44	11.56

Table C.28 (Continued): PFK(NR) Gene (21 Introns) Versus Contrast Genes (Gray Cells Reject Ho at α= 0.01 Level)

Contrast Gene	Intron 12	Intron 13	Intron 14	Intron 15	Intron 16	Intron 17	Intron 18	Intron 19	Intron 20	Intron 21	Mean	SD
ABL(O)											46.30	36.89
BCR(O)	9.09	24.77	8.86	21.42	15.10	39.91	9.71	7.77	6.72	26.63	18.71	11.05
FMS(O)	13.38	22.71	9.58	23.81	10.03	40.76	11.47	13.28	19.65	13.87	29.31	41.19
BCL-3(O)											30.51	18.95
FOS(O)											16.85	7.47
HST-1(O)											16.61	7.50
INT-2(O)											12.97	1.39
LCK(O)											19.65	9.29
MYC(O)											17.49	1.30
L-MYC(O)											16.41	3.17
N-MYC(O)											12.73	0.75
MAX(O)											10.88	NA
PIM1(O)											28.90	14.95
KIT(O)	29.12	47.56	28.57	33.17	31.34	45.78	23.13	32.77	18.77		41.52	21.55
RAFA1(O)	16.84	14.84	22.92	19.13							23.78	22.11
FPS(O)	9.31	15.92	11.39	15.87	8.25	15.86	14.98				14.59	5.02
WNT-1(O)											17.26	5.06
K-RAS2(O)											67.26	60.07
MLH1(TS)	24.52	30.70	18.35	25.09	17.25	41.42	17.63				40.61	20.78
MSH2(TS)	45.51	53.02	15.95	45.31							62.87	31.89
MTS1(TS)											16.27	9.78
RB(TS)	64.66	46.03	16.35	57.73	29.93	74.41	47.20	39.47	48.23	150.85	75.46	51.84
NF1(TS)	45.80	46.86	32.03	44.97	31.70	56.23	37.25	30.14	33.27	88.84	56.93	26.03
G6PD(NR)	8.91										15.88	8.07
PGK(NR)											43.34	20.86
ADOL(NR)											56.97	34.55
GTP(NR)											38.05	30.67
TPI(NR)											19.88	7.23
GCK(NR)											18.39	5.54
GAPDH(NR)											16.01	8.59
SDHB(NR)											36.53	25.82
HKII(NR)	14.89	27.98	11.70	20.20	13.85	28.20					18.51	8.24

Table C.29: TPI(NR) Gene (6 Introns) Versus Contrast Genes (Gray Cells Reject Ho at $\alpha = 0.01$ Level)

Contrast Gene	# Introns in Contrast Gene	# Introns Compared	Intron 1	Intron 2	Intron 3	Intron 4	Intron 5	Intron 6	Mean	SD
ABL(O)	10	6	372.93	29.09	15.94	28.57	22.41	22.61	81.92	142.65
BCR(O)	22	6	210.59	11.02	17.04	18.32	14.27	12.08	47.22	80.08
FMS(O)	21	6	195.41	9.67	20.92	15.85	5.59	19.11	44.43	74.19
BCL-3(O)	8	6	108.28	18.34	43.58	19.45	43.58	32.80	44.34	33.22
FOS(O)	3	3	14.49	9.13	13.39				12.34	2.83
HST-1(O)	2	2	31.47	10.64					21.06	14.73
INT-2(O)	2	2	18.46	17.20					17.83	0.89
LCK(O)	11	6	43.16	9.01	29.44	21.06	60.87	10.42	28.99	20.12
MYC(O)	2	2	30.25	15.56					22.91	10.39
L-MYC(O)	2	2	36.58	12.64					24.61	16.93
N-MYC(O)	2	2	13.66	14.81					14.24	0.81
MAX(O)	1	1	52.1						52.10	NA
PIM1(O)	5	5	9.98	24.01	35.10	25.34	23.69		23.63	8.96
KIT(O)	20	6	27.03	29.80	15.44	50.88	52.49	22.99	33.11	15.19
RAFA1(O)	15	6	58.84	15.06	16.26	16.11	12.13	56.29	29.12	22.10
FPS(O)	18	6	20.41	19.80	18.53	7.18	17.51	13.34	16.13	5.05
WNT-1(O)	3	3	25.23	15.31	23.58				21.37	5.32
K-RAS2(O)	4	4	46.56	38.42	23.63	85.06			48.42	26.21
MLH1(TS)	18	6	57.76	25.58	10.10	41.87	46.59	39.42	36.89	16.77
MSH2(TS)	15	6	24.52	57.34	14.68	72.53	51.39	70.67	48.52	23.97
MTS1(TS)	3	3	56.76	13.31	15.00				28.35	24.61
RB(TS)	26	6	76.99	44.88	19.49	87.62	57.05	72.30	59.72	24.82
NF1(TS)	56	6	57.71	25.93	21.01	32.88	70.82	55.63	44.00	20.11
G6PD(NR)	12	6	26.71	13.01	34.75	12.55	35.20	11.26	22.25	11.35
PGK(NR)	10	6	53.96	25.72	24.08	33.75	25.51	29.14	32.03	11.29
ADOL(NR)	8	6	320.17	22.70	12.67	42.06	29.97	28.88	76.07	119.97
GTP(NR)	8	6	44.29	22.15	13.61	14.68	39.41	8.77	23.82	14.69
PFK(NR)	21	6	14.00	13.39	27.18	28.99	22.38	13.37	19.88	7.23
GCK(NR)	9	6	84.83	10.64	38.38	16.95	28.33	14.52	32.27	27.69
GAPDH(NR)	8	6	26.65	12.89	20.77	14.87	15.00	9.45	16.61	6.15
SDHB(NR)	7	6	60.86	20.52	15.40	21.83	12.66	20.48	25.29	17.78
HKII(NR)	17	6	25.17	12.90	18.05	15.76	21.69	20.19	18.96	4.37

Table C.30: GCK(NR) Gene (9 Introns) Versus Contrast Genes (Gray Cells Reject Ho at $\alpha = 0.01$ Level)

Contrast Gene	# Introns in Contrast Gene	# Introns Compared	Intron 1	Intron 2	Intron 3	Intron 4	Intron 5	Intron 6	Intron 7	Intron 8	Intron 9	Mean	SD
ABL(O)	10	9	331.26	65.76	82.12	37.45	39.84	32.52	27.55	124.1	33.15	85.97	97.20
BCR(O)	22	9	121.53	15.41	27.62	25.77	27.32	15.56	20.81	76.34	30.71	40.12	35.60
FMS(O)	21	9	167.64	29.55	32.60	21.75	28.35	41.74	37.15	11.39	24.71	43.88	47.24
BCL-3(O)	8	8	91.46	40.42	57.18	21.67	28.96	31.82	17.44	130.85		52.48	39.62
FOS(O)	3	3	80.63	17.54	49.10							49.09	31.55
HST-1(O)	2	2	119.10	9.16								64.13	77.74
INT-2(O)	2	2	104.68	17.15								60.91	61.89
LCK(O)	11	9	43.90	6.74	17.89	28.79	41.65	15.13	14.70	49.42	19.14	26.37	15.21
MYC(O)	2	2	138.46	41.74								90.10	68.39
L-MYC(O)	2	2	105.74	34.24								69.99	50.56
N-MYC(O)	2	2	50.45	39.59								45.02	7.68
MAX(O)	1	1	18.70									18.70	NA
PIM1(O)	5	5	48.28	43.07	15.55	49.30	45.30					40.30	14.05
KIT(O)	20	9	44.07	45.90	69.71	72.68	72.96	28.14	26.78	115.86	89.43	62.84	29.39
RAFA1(O)	15	9	65.71	19.86	85.90	16.98	14.15	85.78	7.37	27.68	13.91	37.48	32.23
FPS(O)	18	9	44.05	24.44	41.53	11.21	35.44	6.15	24.71	30.68	27.73	27.33	12.66
WNT-1(O)	3	3	40.13	17.59	27.86							28.53	11.29
K-RAS2(O)	4	4	102.35	80.17	120.77	108.37						102.91	16.99
MLH1(TS)	18	9	80.27	54.30	85.85	59.18	57.90	57.52	36.84	107.25	59.70	66.53	20.94
MSH2(TS)	15	9	58.22	107.13	102.01	81.17	70.17	88.17	56.29	136.99	73.35	85.94	25.97
MTS1(TS)	3	3	50.61	15.79	56.84							41.08	22.13
RB(TS)	26	9	119.70	98.55	122.85	117.94	74.96	103.11	76.57	135.24	93.10	104.67	20.96
NF1(TS)	56	9	62.22	51.93	113.10	50.66	80.59	74.07	49.81	106.30	79.66	74.26	23.45
G6PD(NR)	12	9	37.96	23.62	21.98	26.12	22.58	6.80	13.03	43.86	29.11	25.01	11.36
PGK(NR)	10	9	54.54	50.87	58.22	47.88	43.62	38.08	29.33	83.23	42.94	49.86	15.24
ADOL(NR)	8	8	248.10	38.95	114.80	51.29	52.34	32.50	47.10	172.25		94.67	78.26
GTP(NR)	8	8	31.36	38.60	32.82	23.01	54.94	10.15	28.23	70.03		36.14	18.72
PFK(NR)	21	9	13.02	10.34	20.72	26.82	18.87	15.72	24.88	21.04	14.08	18.39	5.54
TPI(NR)	6	6	84.83	10.64	38.38	16.95	28.33	14.52				32.27	27.69
GAPDH(NR)	8	8	121.12	18.92	9.17	20.25	17.07	9.22	14.37	26.88		29.63	37.43
SDHB(NR)	7	7	69.67	48.46	53.19	40.53	30.98	28.66	27.66			42.74	15.47
HKII(NR)	17	9	26.48	18.59	36.55	30.16	41.70	19.04	16.52	36.55	18.62	27.13	9.50

Table C.31: GAPDH(NR) Gene (8 Introns) Versus Contrast Genes (Gray Cells Reject Ho at α = 0.01 Level)

Contrast Gene	# Introns in Contrast Gene	# Introns Compared	Intron 1	Intron 2	Intron 3	Intron 4	Intron 5	Intron 6	Intron 7	Intron 8	Mean	SD
ABL(O)	10	8	237.46	108.72	17.37	28.63	10.26	21.47	27.44	27.10	59.81	78.17
BCR(O)	22	8	174.22	79.07	5.74	28.36	7.71	15.47	16.96	16.90	43.05	57.92
FMS(O)	21	8	166.89	113.56	7.00	24.72	15.56	34.22	32.60	13.18	50.97	57.67
BCL-3(O)	8	8	115.27	112.19	31.34	26.65	7.25	11.31	22.99	58.00	48.13	43.28
FOS(O)	3	3	22.47	31.97	17.86						24.10	7.19
HST-1(O)	2	2	16.23	30.52							23.38	10.10
INT-2(O)	2	2	25.68	153.06							89.37	90.07
LCK(O)	11	8	62.35	13.32	16.38	21.00	25.48	14.39	13.72	10.27	22.11	16.95
MYC(O)	2	2	30.44	85.04							57.74	38.61
L-MYC(O)	2	2	16.32	97.77							57.05	57.59
N-MYC(O)	2	2	38.9	88.59							63.74	35.14
MAX(O)	1	1	66.4								66.40	NA
PIM1(O)	5	5	8.82	32.93	10.33	25.55	24.48				20.42	10.44
KIT(O)	20	8	31.58	52.65	33.95	37.52	43.49	15.52	22.31	32.57	33.70	11.56
RAFA1(O)	15	8	74.82	29.38	19.04	12.65	11.33	55.92	2.78	22.46	28.55	24.59
FPS(O)	18	8	24.87	26.71	9.50	11.66	10.14	7.66	12.23	23.59	15.79	7.84
WNT-1(O)	3	3	43.85	14.34	17.85						25.35	16.12
K-RAS2(O)	4	4	25.46	135.19	36.63	55.87					63.29	49.56
MLH1(TS)	18	8	41.91	73.26	29.14	33.38	31.83	43.56	32.28	33.36	39.84	14.42
MSH2(TS)	15	8	28.43	129.01	41.90	37.78	36.87	58.32	47.54	42.90	52.84	31.97
MTS1(TS)	3	3	44.69	28.99	12.93						28.87	15.88
RB(TS)	26	8	54	159.96	40.23	55.15	47.08	62.58	59.98	42.27	65.16	39.13
NF1(TS)	56	8	49.64	58.23	36.68	20.66	48.98	55.39	44.18	38.41	44.02	12.06
G6PD(NR)	12	8	46.84	89.38	8.21	19.85	8.04	12.77	6.91	25.53	27.19	28.43
PGK(NR)	10	8	60.5	66.35	31.00	24.01	25.21	34.68	35.98	22.11	37.48	16.84
ADOL(NR)	8	8	273.30	87.16	30.97	32.61	26.91	28.85	37.76	28.00	68.19	85.25
GTP(NR)	8	8	59.50	64.42	11.61	18.58	23.29	8.54	24.43	18.64	28.63	21.30
PFK(NR)	21	8	16.23	8.02	12.74	32.84	6.14	12.47	22.98	16.70	16.01	8.59
TPI(NR)	6	6	26.65	12.89	20.77	14.87	15	9.45			16.61	6.15
GCK(NR)	9	8	121.12	18.92	9.17	20.25	17.07	9.22	14.37	26.88	29.63	37.43
SDHB(NR)	7	7	66.06	54.59	16.42	23.24	9.96	18.99	23.26		30.36	21.22
HKII(NR)	17	8	33.41	20.77	11.09	21.55	15.53	17.87	16.24	23.74	20.03	6.70

144

Table 6.32: SDHB(NR) Gene (7 Introns) Versus Contrast Genes (Gray Cells Reject Ho at α = 0.01 Level)

Contrast Gene	# Introns in Contrast Gene	# Introns Compared	Intron 1	Intron 2	Intron 3	Intron 4	Intron 5	Intron 6	Intron 7	Mean	SD
ABL(O)	10	7	34.08	9.43	38.76	18.87	7.66	9.25	20.40	19.78	12.46
BCR(O)	22	7	19.24	50.95	11.26	47.11	12.04	15.02	23.39	25.57	16.60
FMS(O)	21	7	31.42	33.08	24.05	41.03	13.29	31.38	19.92	27.74	9.27
BCL-3(O)	8	7	50.05	26.78	112.75	38.01	40.13	69.92	30.14	52.54	30.18
FOS(O)	3	3	36.74	20.88	19.80					25.81	9.48
HST-1(O)	2	2	108.4	63.51						85.96	31.74
INT-2(O)	2	2	101.75	67.92						84.84	23.92
LCK(O)	11	7	31.99	34.71	51.40	36.92	59.21	32.77	24.70	38.81	12.09
MYC(O)	2	2	42.25	19.21						30.73	16.29
L-MYC(O)	2	2	93.12	22.53						57.83	49.91
N-MYC(O)	2	2	28.91	20.72						24.82	5.79
MAX(O)	1	1	37.39							37.39	NA
PIM1(O)	5	5	38.53	46.52	46.43	19.13	12.67			32.65	15.80
KIT(O)	20	7	18.95	10.94	27.39	16.95	37.14	21.67	8.22	20.18	9.85
RAFA1(O)	15	7	29.8	21.24	54.21	22.67	14.14	53.57	13.52	29.88	17.30
FPS(O)	18	7	69.51	46.75	27.47	23.50	8.78	23.52	34.89	33.49	19.67
WNT-1(O)	3	3	55.65	59.39	55.16					56.73	2.31
K-RAS2(O)	4	4	49.35	21.60	67.86	44.14				45.74	19.04
MLH1(TS)	18	7	32.63	7.02	29.93	18.40	26.84	28.06	23.91	23.83	8.70
MSH2(TS)	15	7	20.49	31.56	59.52	50.69	38.14	71.55	34.25	43.74	17.71
MTS1(TS)	3	3	29.15	61.49	24.23					38.29	20.24
RB(TS)	26	7	22.96	13.64	57.90	45.82	44.13	60.03	42.02	40.93	17.11
NF1(TS)	56	7	25.92	14.75	66.06	13.97	50.26	43.24	25.81	34.29	19.50
G6PD(NR)	12	7	23.79	37.15	34.97	59.81	34.74	36.21	19.20	35.12	12.89
PGK(NR)	10	7	14.77	10.34	31.21	12.65	21.49	24.20	32.33	21.00	8.81
ADOL(NR)	8	7	41.86	30.15	34.96	21.23	18.58	11.3	22.25	25.76	10.46
GTP(NR)	8	7	15.91	16.91	10.40	23.77	16.04	14.67	13.93	15.95	4.06
PFK(NR)	21	7	19.06	23.98	23.74	63.49	16.31	26.36	82.74	36.53	25.82
TPI(NR)	6	6	60.86	20.52	15.4	21.83	12.66	20.48		25.29	17.78
GCK(NR)	9	7	69.67	48.46	53.19	40.53	30.98	28.66	27.66	42.74	15.47
GAPDH(NR)	8	7	66.06	54.59	16.42	23.24	9.96	18.99	23.26	30.36	21.22
HKII(NR)	17	7	26.88	18.93	21.33	13.61	26.43	29.67	23.84	22.96	5.47

Table C.33: HKII(O) Gene (17 Introns) Versus Contrast Genes (Gray Cells Reject Ho at α= 0.01 Level).

Contrast Gene	# Introns in Contrast Gene	# Introns Compared	Intron 1	Intron 2	Intron 3	Intron 4	Intron 5	Intron 6	Intron 7	Intron 8	Intron 9	Intron 10	Intron 11
ABL(O)	10	10	40.31	12.88	16.58	26.1	30.15	41.97	17.45	36.24	8.98	20.17	
BCR(O)	22	17	27.38	24.68	12.57	21.87	30.71	28.94	13.42	19.70	6.34	23.22	11.23
FMS(O)	21	17	43.07	15.55	7.32	17.18	23.83	36.58	41.76	14.68	7.56	34.71	23.51
BCL-3(O)	8	8	40.52	18.41	63.85	32.55	29.93	53.27	18.48	85.14			
FOS(O)	3	3	22.06	10.00	17.59								
HST-1(O)	2	2	38.73	29.52									
INT-2(O)	2	2	26.24	48.63									
LCK(O)	11	11	38.53	16.08	25.37	22.39	44.87	29.31	20.47	21.82	7.12	11.26	31.41
MYC(O)	2	2	27.86	8.43									
L-MYC(O)	2	2	35.14	6.43									
N-MYC(O)	2	2	20.88	7.47									
MAX(O)	1	1	24.49										
PIM1(O)	5	5	26.66	37.24	29.67	25.93	33.11						
KIT(O)	20	17	28.00	16.79	25.83	37.24	52.88	29.46	14.44	58.78	46.83	16.46	47.10
RAFA1(O)	15	15	24.67	17.70	15.51	9.84	17.55	85.57	8.19	22.64	4.42	13.18	18.91
FPS(O)	18	17	26.40	25.30	9.13	14.63	27.19	20.20	10.49	36.79	27.87	24.49	12.72
WNT-1(O)	3	3	24.34	22.69	30.38								
K-RAS2(O)	4	4	39.45	29.74	57.08	66.10							
MLH1(TS)	18	17	31.67	15.73	26.62	30.15	52.77	57.90	30.95	42.36	24.15	21.59	48.78
MSH2(TS)	15	15	16.00	41.70	41.78	66.23	61.01	73.82	36.48	67.54	42.11	28.03	76.14
MTS1(TS)	3	3	34.97	25.57	13.24								
RB(TS)	26	17	32.56	27.31	44.86	65.90	79.03	106.36	58.03	66.77	59.35	46.69	70.51
NF1(TS)	56	17	31.56	20.18	49.67	26.60	83.04	64.34	35.30	66.09	49.12	27.43	88.06
G6PD(NR)	12	12	24.93	15.71	16.09	24.58	46.26	22.96	19.66	24.82	8.65	21.38	15.56
PGK(NR)	10	10	25.32	11.69	21.78	21.21	26.73	47.07	27.90	28.26	34.15	24.78	
ADOL(NR)	8	8	45.58	17.81	21.33	33.37	47.50	39.53	41.59	49.01			
GTP(NR)	8	8	26.28	11.90	15.00	23.43	46.25	20.36	22.84	34.77			
PFK(NR)	21	17	16.83	11.77	12.99	34.47	17.85	11.77	24.16	23.24	3.79	29.44	11.56
TPI(NR)	6	6	25.17	12.90	18.05	15.76	21.69	20.19					
GCK(NR)	9	9	26.48	18.59	36.55	30.16	41.70	19.04	16.52	36.55	18.62		
GAPDH(NR)	8	8	33.41	20.77	11.09	21.55	15.53	17.87	16.24	23.74			
SDHB(NR)	7	7	26.88	18.93	21.33	13.61	26.43	29.67	23.84				

Table C.33 (Continued): HKII(O) Gene (17 Introns) Versus Contrast Genes (Gray Cells Reject Ho at α= 0.01 Level)

Contrast Gene	Intron 12	Intron 13	Intron 14	Intron 15	Intron 16	Intron 17	Mean	SD
ABL(O)							25.08	11.70
BCR(O)	13.07	29.16	13.31	7.13	6.39	23.63	18.40	8.46
FMS(O)	13.08	29.26	17.02	7.21	11.36	31.53	22.07	12.09
BCL-3(O)							42.77	23.32
FOS(O)							16.55	6.10
HST-1(O)							34.13	6.51
INT-2(O)							37.44	15.83
LCK(O)							24.42	11.22
MYC(O)							18.15	13.74
L-MYC(O)							20.79	20.30
N-MYC(O)							14.18	9.48
MAX(O)							24.49	NA
PIM1(O)							30.52	4.71
KIT(O)	12.50	26.98	30.61	16.37	37.17	8.78	29.78	15.01
RAFA1(O)	17.00	26.74	33.76	33.64			23.29	19.20
FPS(O)	11.71	35.97	9.69	31.22	12.63	35.25	21.86	9.87
WNT-1(O)							25.81	4.05
K-RAS2(O)							48.09	16.50
MLH1(TS)	7.00	20.25	18.54	10.65	8.29	20.25	27.51	15.24
MSH2(TS)	15.95	26.84	22.98	27.89			42.97	20.93
MTS1(TS)							24.60	10.90
RB(TS)	26.93	29.48	25.20	52.76	35.83	31.45	50.53	22.50
NF1(TS)	24.24	35.76	42.22	40.65	29.45	17.00	42.98	21.21
G6PD(NR)	21.13						21.81	9.11
PGK(NR)							26.89	9.18
ADOL(NR)							36.97	11.85
GTP(NR)							25.11	11.01
PFK(NR)	14.89	27.98	11.70	20.20	13.85	28.20	18.51	8.24
TPI(NR)							18.96	4.37
GCK(NR)							27.13	9.50
GAPDH(NR)							20.03	6.70
SDHB(NR)							22.96	5.47

APPENDIX D

SAS Code

Macro to Analyze Number of Introns, Means of Intronic Log Lengths and Standard Deviations of Intronic Log Lengths in a Gene

```
title 'Intron Attributes in Contrast Genes';
data genepool;
input gene & $12. intnum intmn intsd lnintmn lnintsd@@;
cards;
```

ABL(O)	10	22724.3	62361.11	7.957296	1.837927
BCR(O)	22	5943.0	15089.64	7.530286	1.304262
FMS(O)	21	2622.095	5632.928	6.676911	1.572582
BCL-3(O)	8	1218.625	1676.15	6.408406	1.237932
FOS(O)	3	432.6667	319.5033	5.8088	0.969894
HST-1(O)	2	577.5	55.86144	6.3564	0.096874
INT-2(O)	2	3968.5	2375.172	8.1875	0.638659
LCK(O)	11	986.0909	1867.0	5.710842	1.401179
MYC(O)	2	1500.0	175.3625	7.30975	0.117168
L-MYC(O)	2	1667.5	1843.427	6.94695	1.484571
N-MYC(O)	2	1765.0	1231.78	7.33635	0.764595
MAX(O)	1	483.0	.	6.18	.
PIM1(O)	5	514.4	622.4386	5.565122	1.311433
KIT(O)	20	1574.167	1408.351	6.788102	1.296034
RAFA1(O)	15	554.0	583.6962	5.69871	1.187702
FPS(O)	18	473.5556	531.8399	5.680134	0.973222
WNT-1(O)	3	625.6667	141.8462	6.419653	0.246156
K-RAS2(O)	4	8979.75	6414.456	8.796863	1.021532
MLH1(TS)	18	3069.667	2774.064	7.654244	1.020609
MSH2(TS)	15	4573.333	4755.218	8.053468	0.836118
MTS1(TS)	3	536.6667	266.5827	6.170747	0.632422
RB(TS)	26	6644.962	14556.52	7.754034	1.459014
NF1(TS)	56	4285.979	13155.08	6.99366	1.412317
G6PD(NR)	12	1102.5	2765.385	5.749208	1.33521
PGK(NR)	10	2092.1	2110.976	6.998464	1.293345
ADOL(NR)	8	1581.75	1516.879	7.05172	0.798806
GTP(NR)	8	315.0	148.8057	5.607898	0.634424
PFK(NR)	21	1182.857	1915.214	6.406921	1.139428
TPI(NR)	6	341.1667	413.5773	5.370938	0.985794
GCK(NR)	9	2534.667	2694.33	7.486829	0.8163
GAPDH(NR)	8	321.5	533.1883	5.145963	0.982632
SDHB(NR)	7	4950.0	3979.426	8.159084	0.981201
HKII(NR)	17	2360.059	3918.626	6.737161	1.504913

```
;

title 'Cluster Analysis Based on Number of Introns, Mean and SD of Log
Lengths for 33 Contast Genes';

%macro analyze(method,ncl);
proc cluster data=genepool out=tree method=&method std ccc pseudo;
  var intnum lnintmn lnintsd;
  id gene;
  title2;
%let k=1;
%let n=%scan(&ncl,&k);
```

```
%do %while(&n^=);
proc tree data=tree noprint out=out ncl=&n;
copy intnum intmn;
proc plot;
plot intnum*intmn=cluster / hpos=86 vpos=26;
title2 "Plot of &n Clusters from METHOD=&METHOD";
run;
%let k=%eval(&k+1);
%let n=%scan(&ncl,&k);
%end;
%mend;

%analyze(average,)
%analyze(centroid,)
%analyze(complete,)
%analyze(eml,)
%analyze(flexible,)
%analyze(mcquitty,)
%analyze(median,)
%analyze(single,)
%analyze(ward,)
```

SAS Code to Perform Cluster Analysis Based on Mean Chi-square Homology Statistics After Comparing Introns Between Genes

```
options center ls=80;
data homology(type=distance);
infile 'c:\msoffice\excel\frdecl.txt' delimiter='@' missover lrecl=330;
input a b c d e f g h i j k l m n o p q r s t u v w x y z aa ab ac ad ae
af ag genname $ ;
run;
proc cluster data=homology method=average pseudo nosquare ;
id genname;
run;

proc cluster data=homology method=centroid pseudo nosquare ;
id genname;
run;

proc cluster data=homology method=single nosquare ;
id genname;
run;

proc cluster data=homology method=complete nosquare ;
id genname;
run;

proc cluster data=homology method=flexible  nosquare ;
id genname;
run;

proc cluster data=homology method=ward pseudo nosquare   ;
id genname;
run;

proc cluster data=homology method=Mcquitty nosquare   ;
id genname;
run;

proc cluster data=homology method=median pseudo nosquare   ;
id genname;
run;
```

APPENDIX E

S-PLUS CODE

S-Plus Code to Create a Markov Chain Once the DNA Sequence Has Been Read Into 'xtest' from a Text File (Code is for a Chain 50 bp long)

```
>x0_xtest[-50]
>x1_xtest[-1]
rbind[x0,x1]
```

/*The above code is more efficient than that traditional long way (an iterative loop). The only problem is that there is a limit on how much code the rbind procedure can handle. A DNA seqence longer than 10,000 bp must be broken apart in order to use the above procedure. In this case, the user may prefer to create the chain using the iterative loop (even though it will be fairly slow). The code for the long way is also provided*/

```
>M_matrix(0,4,4)
>for (i in 1:49) {
+a_xtest[i]
+b_xtest[i+1]
+M[a,b]_M[a,b]+1}
```

S-Plus Function to Test for Independence Versus a Markov Process

```
> mark.test
function(x)
{
        zij <- matrix(0, ncol = 4, nrow = 4)
        for(i in 1:4) {
                for(j in 1:4) {
                        zij[i, j] <- (((x[i, j]) - ((
                                sum(x[i,  ])) * ((sum(x[
                                , j]))/(sum(x)))))^2)/((
                                sum(x[i,  ])) * ((sum(x[
                                , j]))/(sum(x))))
                        sum(zij)
                }
        }
}
```

S-Plus Function to Compute Chi-square Test Statistics from a Test for Homology Between Intron x and Intron y

```
> homl.test
function(x, y)
{
        zij <- matrix(0, ncol = 4, nrow = 4)
        for(i in 1:4) {
                for(j in 1:4) {
                        zij[i, j] <- (((sum(x[i,  ])) * (sum(y[i,  ])))/
                                ((x[i, j]) + (y[i, j]))) * (((((x[i, j])/
                                (sum(x[i,  ]))) - ((y[i, j])/(sum(y[i,
                                ])))))^2)
                }
        }
        sum(zij)
}
>
```